The American Economic Review

ARTICLES

JUNE 1990

THE AMERICAN ECONOMIC ASSOCIATION

• Printed at Banta Company, Menasha, Wisconsin.

• No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002-8282), March 1990, Vol. 80, No. 2, is published five times a year (March, May, June, September, December) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: $42.00, $50.40, or $58.80 depending on income. A membership also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add $16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 2014 Broadway, Suite 305, Nashville, TN 37203.

Founded in 1885

Officers

President
GERARD DEBREU
University of California-Berkeley

President-elect
THOMAS C. SCHELLING
Harvard University

Vice-Presidents
RUDIGER W. DORNBUSCH
Massachusetts Institute of Technology
ALLAN H. MELTZER
Carnegie Mellon University

Secretary-Treasurer
C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review
ORLEY C. ASHENFELTER
Princeton University

Editor of The Journal of Economic Literature
JOHN PENCAVEL
Stanford University

Editor of The Journal of Economic Perspectives
JOSEPH E. STIGLITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee
GEORGE A. AKERLOF
University of California-Berkeley
ISABEL V. SAWHILL
Urban Institute
STANLEY FISCHER
World Bank
LAWRENCE H. SUMMERS
Harvard University
GREGORY C. CHOW
Princeton University
SUSAN ROSE-ACKERMAN
Yale University

EX OFFICIO Member
ROBERT EISNER
Northwestern University

JACOB MINCER

DISTINGUISHED FELLOW

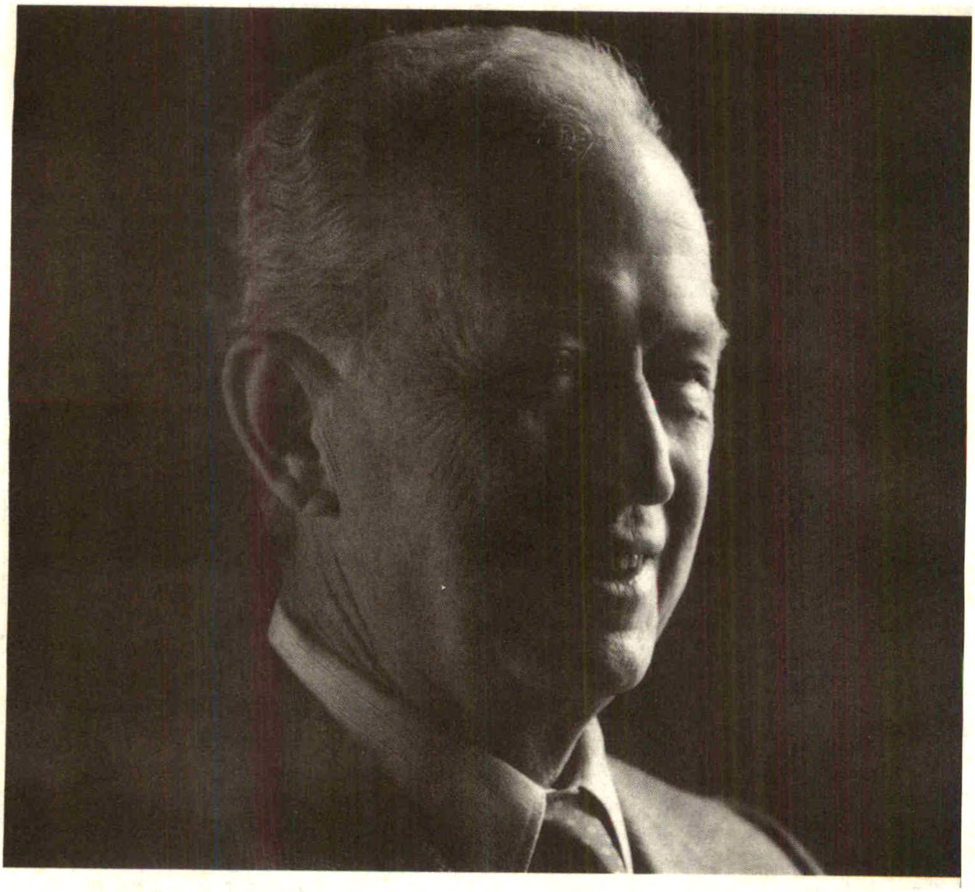
1988

Jacob Mincer's research reveals a rare combination of imaginative empirical analysis guided by a command of theory. His work and professional style have set the standard of empirical economics, especially in the field of labor economics, where he has made major contributions to the understanding of the determinants of earnings and the labor force participation of married women.

His book *Schooling, Experience and Earnings* is a classic that has enormously influenced the quantitative study of earnings in countries throughout the world. He based his analysis on the theory of investment in human capital, a framework that continues to motivate most empirical earnings studies. He showed that a few variables—especially education and measures of labor force experience—explain a significant fraction of differences in earnings among males.

He has subsequently extended this analysis to consider the earnings of women, job separations, differences between the effects on earnings of job tenure and general employment experience, and other variables.

Mincer was also among the first to examine the labor force participation of married women in a model that highlights the relation between the decisions of husbands and wives, and the effect of family size on labor force participation. His approach continues to be followed and extended as social scientists study further the remarkable growth in participation of married women in all developed and many developing countries.



Jacob Mincer

THE AMERICAN ECONOMIC REVIEW

June 1990

VOLUME 80, NUMBER 3

Articles

Shorter Papers

Erratum

•Submit manuscripts (4 copies), 50 pages maximum, single-sided, double-spaced, to:

Orley Ashenfelter, Editor, *AER*, 209 Nassau Street, Princeton, NJ 08542–4607.

•Submission fee: $50 for members; $100 for nonmembers. Please pay with a check or money order payable in United States Dollars. Canadian and Foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Style guides will be provided upon request.

Editorial Statement

It is the policy of the *American Economic Review* to publish papers only where the data used in the analysis are clearly and precisely documented, are readily available to any researcher for purposes of replication, and where details of the computations sufficient to permit replication are provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary, or if, for some other reason, the above requirements cannot be met.

Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records

By JOSHUA D. ANGRIST*

The randomly assigned risk of induction generated by the draft lottery is used to construct estimates of the effect of veteran status on civilian earnings. These estimates are not biased by the fact that certain types of men are more likely than others to service in the military. Social Security administrative records indicate that in the early 1980s, long after their service in Vietnam was ended, the earnings of white veterans were approximately 15 percent less than the earnings of comparable nonveterans. (JEL 824)

A central question in the debate over military manpower policy is whether veterans are adequately compensated for their service. The political process clearly reflects the desire to compensate veterans: since World War II, millions of veterans have enjoyed benefits for medical care, education and training, housing, insurance, and job placement. Recent legislation provides additional benefits for veterans of the Vietnam era. Yet, academic research has not shown conclusively that Vietnam (or other) veterans are worse off economically than nonveterans. Many studies find that Vietnam veterans earn less than nonveterans, but others find positive effects, or effects that vary with age and schooling. Regarding the general position of veterans, a member of the Twentieth Century Fund's Task Force on Policies Toward Veterans concludes that "Within any age group, veterans have higher incomes, more education, and lower unemployment rates than their nonveteran counterparts."[1]

The goal of this paper is to measure the long-term labor market consequences of military service during the Vietnam era. Previous research comparing civilian earnings by veteran status may be biased by the fact that certain types of men are more likely to serve in the armed forces than others. For example, men with relatively few civilian opportunities are probably more likely to enlist. Estimation strategies that do not control for differences in civilian earnings potential will incorrectly attribute lower civilian earnings of veterans to military service. The research reported here overcomes such statistical problems by using the Vietnam era draft

*Department of Economics, Harvard University, Cambridge, MA 02138. Grateful thanks go to Warren Buckler, Cresston Smith, Ada Enis, and Bea Matsui for their assistance in producing the Social Security data; to Chester Bowie for his help in producing the SIPP data; and to Mike Dove for providing DMDC administrative records. Special thanks also go to David Card and Whitney Newey, from whose instruction and comments I have benefited greatly, and to Alan Krueger and an anonymous referee, whose careful reviews of an earlier draft led to substantial improvement. Data collection for this project was funded by the Princeton Industrial Relations Section. Funds for computation and financial support of the author were provided by the Industrial Relations Section, the Princeton Department of Economics, the Sloan Foundation, and the Olin Foundation.

[1] The quote is from Michael Taussig (1974, p. 51). Legislation pertaining to veterans benefits is outlined in Veterans Administration (1984) and in other annual reports of the Veterans Administration. Studies by Sherwin Rosen and Paul Taubman (1982), Saul Schwartz (1986), and Jon Crane and David Wise (1987) find that Vietnam veterans earn less than nonveterans. Dennis DeTray (1982) and Mark Berger and Barry Hirsch (1983) find some positive effects for different age and schooling classes, and Veterans Administration (1981a) researchers find an overall positive effect.

lotteries to set up a natural experiment that randomly influenced who served in the military.[2]

Section I describes the Social Security administrative records used in the empirical work and provides background on the draft lotteries. In each lottery, priority for induction was determined by a Random Sequence Number (RSN) from 1–365 that was assigned to birthdates in the cohort being drafted. Men were called for induction by RSN up to a ceiling determined by the Defense Department, and only men with lottery numbers below the ceiling could have been drafted. Therefore, men with lottery numbers below the ceiling are referred to here as "draft-eligible."

The empirical analysis begins in Section II with estimates of the effect of draft eligibility on earnings. If draft eligibility is correlated with veteran status but uncorrelated with other variables related to earnings, then earnings differences by draft-eligibility status can be attributed to military service. In Section III, information on the proportions of draft-eligible and draft-ineligible men who actually served in the military is used to convert estimates of the effect of draft eligibility into estimates of the effect of military service. The assumptions underlying this procedure are those that justify instrumental variables estimation; in principle, any function of the RSN provides a legitimate instrument for veteran status. In the second part of Section III, an instrumental variables estimation strategy is developed which is more efficient than one based solely on draft-eligibility status. Results in Section III indicate that white veterans earn approximately 15 percent less than nonveterans as much as ten years after their discharge from the military.

Section IV tests the hypothesis that veterans earn less than nonveterans because they have less civilian labor market experience. Results in this section suggest that the earnings loss to white veterans is equivalent to a loss of two years of civilian labor market experience. Section V reviews some of the potential pitfalls in estimation based on the draft lottery. Section VI offers conclusions and indicates directions for future research.

I. Background and Data

A. *National Random Selection*[3]

There were five draft lotteries during the Vietnam War period. The 1970 lottery covered 19- to 26-year-old men born in 1944–50, although most of the men drafted in 1970 were born in 1950. Other lotteries were restricted to 19- and 20-year-olds. The 1971 lottery covered men born in 1951, the 1972 lottery covered men born in 1952, and so on, through 1975. However, no one was drafted after 1972, and congressional conscription authority expired in July 1973.

Draft lottery RSNs were randomly assigned in a televised drawing held a few months before men reaching draft age were to be called.[4] Draft-eligibility ceilings—RSN 195 in 1970, RSN 125 in 1971, and RSN 95 in 1972—were announced later in the year, once Defense Department manpower needs were known. As a consequence of this delay, many men with low numbers volunteered for the military to avoid being drafted and to improve their terms of service (Angrist 1989b). There was even a behavioral response to the lottery in enlistment rates for the 1953 cohort, although no one born in 1953 was drafted. In the analysis that follows, the "draft-eligibility ceiling" for men born in 1953 is set at RSN 95, the highest lottery number called in 1972.

[2] A candid assessment of the problems caused by nonrandom selection for military service is given by Crane and Wise (1987), who note they were unable to use econometric sample selection models to generate robust estimates of the effects of military service on civilian earnings. The first researchers to use the lottery to solve the selection problem were Norman Hearst, Tom Newman, and Stephen Hulley (1986), who present lottery-based estimates of delayed effects of military service on mortality.

[3] This section draws on Curtis Tarr (1981) and the Selective Service System (1986).

[4] Men born from 1944–49 were already of draft age when the 1970 lottery was held on December 1, 1969. For nonveterans in this group, subsequent liability for service was determined by 1970 lottery numbers.

Only the initial selection process was based on RSN order. Subsequent selection from the draft-eligible, nondeferred pool was based on a number of criteria. The most important screening criteria were the pre-induction physical examination and a mental aptitude test. In 1970, for example, half of all registrants failed pre-induction examinations and 20 percent of those who passed were eliminated by physical inspections conducted at induction (Selective Service System, 1971). Of course, the fact that armed forces selection criteria were ultimately not random does not mean that the initial *priority* for induction was not randomly assigned by RSN.

The year 1970 was the last time men over the age of 20 were drafted. In principle, nonveterans born between 1944 and 1949 continued to be at risk of induction in the 1970 lottery, but the majority of men who ended up serving from these cohorts had already entered the military by the time of the 1970 lottery drawing. Veterans born from 1944–49 who managed to avoid service until 1970 may not constitute a representative sample. Therefore, the analysis here is restricted to men who turned 19 in the year they were at risk of induction. This sample includes men who were born between 1950 and 1953.

B. *Social Security Earnings Data*

Earnings data used in this study are drawn from the Social Security Administration's (SSA) Continuous Work History Sample (CWHS). The CWHS data set, described in detail in the Appendix, is a one percent sample drawn from all possible Social Security numbers. The CWHS includes two earnings series: the first contains information on the 1964–84 earnings of men in employment covered by FICA (Social Security) up to the Social Security taxable maximum. It also includes FICA taxable earnings from self-employment. The second series, beginning in 1978, contains total compensation as reported on Internal Revenue Service Form W-2, excluding earnings from self-employment. In principle, the W-2 earnings data are neither censored nor limited to earnings from Social Security taxable employment. However, because SSA procedures for the collection of W-2 forms are relatively new, W-2 earnings data are probably less reliable than the FICA data.

The original CWHS data set does not contain information on date of birth. SSA programmers matched date of birth variables to the CWHS in a special extract created for this project. Lottery numbers were then matched to dates of birth, using tables published in the 1969–73 Semiannual Reports of the Director of Selective Service.

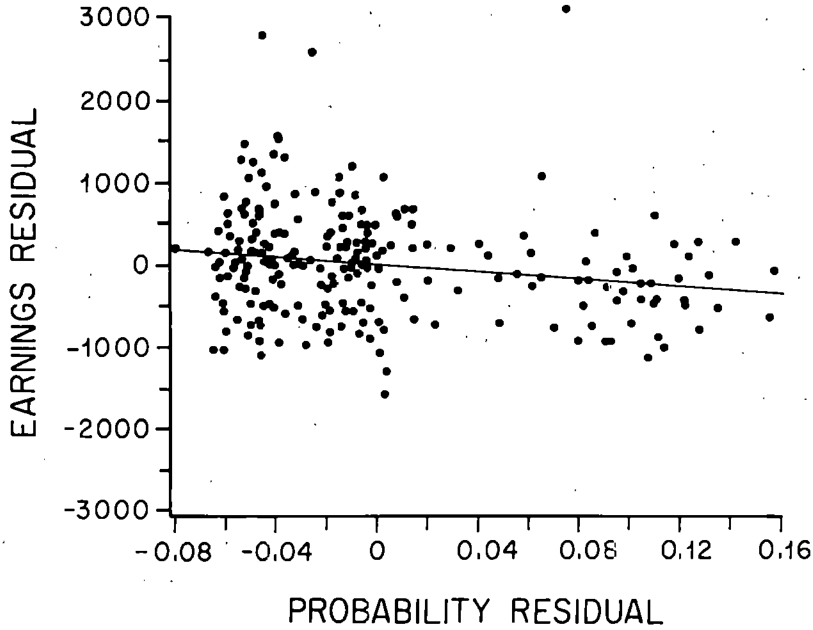
The Internal Revenue Service limits disclosure of data collected for tax purposes. To adhere to these disclosure requirements, the SSA could release only aggregate data. The aggregate data set contains sample statistics for cells defined by year of earnings, year of birth, race, and five consecutive lottery numbers. Cell statistics include means, variances, fraction with earnings equal to the taxable maximum, fraction with earnings above the taxable maximum, fraction with zero earnings, and number of observations in each cell.

II. The Effect of Draft Eligibility on Earnings

Figure 1 shows the history of FICA taxable earnings for draft lottery participants born between 1950 and 1953.[5] For each cohort there are two lines drawn: one for draft-eligible men, and one for men with lottery numbers that exempted them from the draft.

The impact of draft eligibility on the earnings profiles is striking. There appears to be no difference in earnings until the year of conscription risk in the draft lottery. Subsequently, the earnings of draft-eligible white men born in 1950–52 fall below the earnings of draft-ineligible white men born in 1950–52. The earnings of draft-eligible nonwhites also fall below the earnings of other non-

[5] Earnings are in 1978 dollars. The deflator used for all tabulations is the CPI on p. 313 of *The Economic Report of the President* (Council of Economic Advisors, 1988).



Notes: The figure plots the history of FICA taxable earnings for the four cohorts born 1950–53. For each cohort, separate lines are drawn for draft-eligible and draft-ineligible men. Plotted points show average real (1978) earnings of working men born in 1953, real earnings + $3000 for men born in 1950, real earnings + $2000 for men born in 1951, and real earnings + $1000 for men born in 1952.

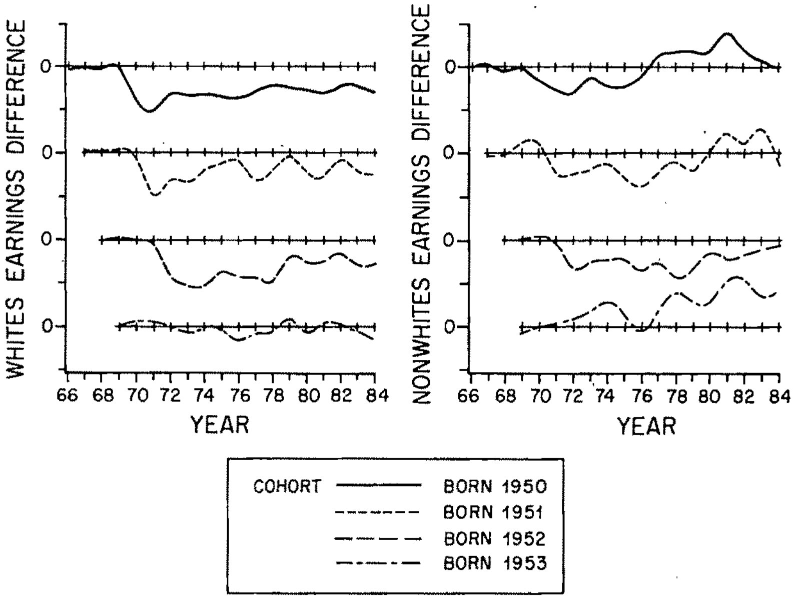
FIGURE 1. SOCIAL SECURITY EARNINGS PROFILES BY DRAFT-ELIGIBILITY STATUS

whites, but the gap appears to narrow and become positive for some nonwhite cohorts in later years. The fact that earnings do not differ by draft-eligibility status before the lotteries is a consequence of the random assignment of draft eligibility. The only thing that distinguishes draft-eligible men from draft-ineligible men is the higher conscription risk faced by eligible men after the lottery.

Figure 2 presents a magnified view of the effect of draft eligibility on earnings. This figure plots the time-series of differences in earnings by draft-eligibility status for each cohort. As in Figure 1, Figure 2 shows no difference during the years before the year of conscription risk, while in subsequent years, the earnings histories diverge. Figure 2 also shows that the loss of earnings to draft-eligible white men was largest during the period they were most likely to be in the service. However, the earnings of draft-eligible white men continued to lag behind the earnings of draft-ineligible white men through 1984.

The picture for nonwhites is less clear. The earnings of draft-eligible nonwhites born in 1950 and 1951 exceed those of draft-ineligible nonwhites in some of the later years. On the other hand, for nonwhites born in 1952, time-series variation in earnings differences by eligibility status is similar to that of whites. The general impression for the three older cohorts of nonwhites is that the earnings of draft-eligible men at least had caught up with the earnings of draft-ineligible men by 1984.

Earnings of white men born in 1953 do not appear to differ by draft-eligibility status. The earnings of draft-eligible nonwhites born in 1953 generally exceed the earnings of nonwhites who were not draft-eligible. Differences between the effect of draft eligi-



Notes: The figure plots the difference in FICA taxable earnings by draft-eligibility status for the four cohorts born 1950–53. Each tick on the vertical axis represents $500 real (1978) dollars.

FIGURE 2. THE DIFFERENCE IN EARNINGS BY DRAFT-ELIGIBILITY STATUS

bility on men born in 1953 and the effect on the three older cohorts might be explained by the transition to an All-Volunteer Force in 1973. Men who volunteer for the military are probably less likely than draftees to suffer a career disadvantage from their service.

Estimates of the effect of draft eligibility are reported in Table 1 for both FICA earnings and W-2 earnings. Standard errors associated with the estimates are reported in parentheses. The statistics in Table 1 show that the loss in FICA earnings to draft-eligible white men is sometimes statistically significant and amounts to 2–3 percent of earnings. Estimated W-2 earnings losses are similar, but tend to be larger and more variable than the estimated losses in FICA earnings. In contrast, differences in earnings by draft-eligibility status for nonwhites rarely exceed their standard errors.

Elsewhere (Angrist 1989c), I have shown that draft-eligible white men are less likely to have earnings above the FICA taxable maximum than draft-ineligible white men. The effect of draft eligibility on nonwhites' probability of being at the taxable maximum, although imprecisely measured, also appears to go in the same direction as the effect of draft eligibility on mean earnings. These results are worth noting because, when the effect of draft eligibility on the probability of being censored has the same sign as the effect on earnings, estimates tabulated using censored data tend to underestimate the true effect.[6]

[6] The effect of censoring on estimated treatment effects is discussed in the appendix. Angrist (1989c) also reports estimates of the effect of draft eligibility on the probability of having no recorded earnings. These tabulations indicate that draft-eligible whites were somewhat more likely to have had FICA earnings during the years in which they were in the service, and that draft-eligible nonwhites are more likely to have had no earnings in recent years. There is no statistically significant evidence for either race, however, of any lasting effect of draft eligibility on the probability of having zero earnings.

Table 1—Draft-Eligibility Treatment Effects for Earnings

	Whites							
	FICA Taxable Earnings				Total W-2 Compensation			
Year	1950	1951	1952	1953	1950	1951	1052	1953
66	−21.8							
	(14.9)							
67	−8.0	13.1						
	(18.2)	(16.4)						
68	−14.9	12.3	−8.9					
	(24.2)	(19.5)	(19.2)					
69	−2.0	18.7	11.4	−4.0				
	(34.5)	(26.4)	(22.7)	(18.3)				
70	−233.8	−44.8	−5.0	32.9				
	(39.7)	(36.7)	(29.3)	(24.2)				
71	−325.9	−298.2	−29.4	27.6				
	(46.6)	(41.7)	(40.2)	(30.3)				
72	−203.5	−197.4	−261.6	2.1				
	(55.4)	(51.1)	(46.8)	(42.9)				
73	−226.6	−228.8	−357.7	−56.5				
	(67.8)	(61.6)	(56.2)	(54.8)				
74	−243.0	−155.4	−402.7	−15.0				
	(81.4)	(75.3)	(68.3)	(68.1)				
75	−295.2	−99.2	−304.5	−28.3				
	(94.4)	(89.7)	(85.0)	(79.6)				
76	−314.2	−86.8	−370.7	−145.5				
	(106.6)	(102.9)	(98.2)	(93.0)				
77	−262.6	−274.2	−396.9	−85.5				
	(117.9)	(112.2)	(111.1)	(107.1)				
78	−205.3	−203.8	−467.1	−65.3	1,059.3	233.2	175.3	−1,974.5
	(132.7)	(127.0)	(127.3)	(123.1)	(2,159.3)	(1,609.4)	(1,567.9)	(912.1)
79	−263.6	−60.5	−236.8	89.2	−1,588.7	523.6	−580.8	−557.9
	(160.5)	(152.3)	(153.9)	(148.7)	(1,575.6)	(1,590.5)	(736.7)	(750.1)
80	−339.1	−267.9	−312.1	−93.8	−1,028.1	85.6	−581.3	−428.7
	(183.2)	(175.3)	(178.2)	(170.7)	(756.8)	(599.8)	(309.1)	(341.5)
81	−435.8	−358.3	−342.8	34.3	−589.6	−71.6	−440.5	−109.5
	(210.5)	(203.6)	(206.8)	(199.0)	(299.4)	(423.4)	(265.0)	(245.2)
82	−320.2	−117.3	−235.1	29.4	−305.5	−72.7	−514.7	18.7
	(235.8)	(229.1)	(232.3)	(222.6)	(345.4)	(372.1)	(296.5)	(281.9)
83	−349.5	−314.0	−437.7	−96.3	−512.9	−896.5	−915.7	30.1
	(261.6)	(253.2)	(257.5)	(248.7)	(441.2)	(426.3)	(395.2)	(318.1)
84	−484.3	−398.4	−436.0	−228.6	−1,143.3	−809.1	−767.2	−164.2
	(286.8)	(279.2)	(281.9)	(272.2)	(492.2)	(380.9)	(376.0)	(366.0)

III. The Effect of Military Service on Earnings

A. *Estimates Using Draft Eligibility*

Estimates of the effects of military service are based on a simple linear model for earnings. Denote the earnings of man i in cohort c at time t by y_{cti}, and let s_i be an indicator of veteran status. Then we may write

$$y_{cti} = \beta_c + \delta_t + s_i \alpha + u_{it}, \tag{1}$$

where β_c is a cohort effect, δ_t is a period effect common to all cohorts, and u_{it} is a residual. The coefficient α is the effect of military service on civilian earnings. If s_i is correlated with the unobserved components of the earnings equation, then α will not be consistently estimated by Ordinary Least Squares (OLS). For example, correlation between s_i and u_{it} may arise because the armed forces' eligibility criteria are correlated with earnings, but not accounted for by the econometrician, or because veterans are

TABLE 1—CONTINUED

Nonwhites

	FICA Taxable Earnings				Total W-2 Compensation			
Year	1950	1951	1952	1953	1950	1951	1952	1953
66	−11.8							
	(27.6)							
67	12.9	−4.0						
	(34.2)	(30.6)						
68	−29.5	−6.2	−12.0					
	(44.5)	(37.3)	(35.0)					
69	−5.1	67.8	3.4	−42.4				
	(66.8)	(53.4)	(43.4)	(36.4)				
70	−99.8	62.2	24.7	−9.0				
	(78.5)	(75.7)	(62.2)	(44.9)				
71	−164.8	−144.3	−25.0	18.2				
	(92.7)	(86.4)	(85.1)	(60.7)				
72	−188.8	−156.7	−208.2	60.4				
	(113.6)	(105.7)	(104.2)	(92.8)				
73	−85.7	−134.8	−175.6	115.5				
	(137.7)	(127.0)	(129.0)	(119.4)				
74	−179.3	−96.7	−181.4	216.5				
	(165.0)	(160.1)	(155.6)	(145.1)				
75	−190.3	−236.1	−183.7	111.6				
	(189.3)	(186.8)	(185.8)	(166.9)				
76	−105.3	−333.7	−308.9	−46.4				
	(214.7)	(215.4)	(216.5)	(199.3)				
77	112.4	−206.8	−251.1	153.5				
	(238.5)	(240.4)	(248.5)	(233.5)				
78	163.6	−108.6	−424.9	381.9	−1,145.0	2,978.2	−4,676.2	−482.7
	(272.6)	(269.2)	(279.4)	(275.7)	(2,395.6)	(2,869.6)	(1,393.1)	(2,206.0)
79	187.0	−210.3	−391.7	312.0	4,005.4	1,545.0	−494.7	−1,043.3
	(317.2)	(323.0)	(324.8)	(326.3)	(2,721.2)	(2,191.1)	(2,683.8)	(1,660.2)
80	203.2	4.8	−212.6	344.0	790.2	376.4	−292.7	288.6
	(363.1)	(368.4)	(372.5)	(370.3)	(648.1)	(533.6)	(440.9)	(416.4)
81	534.5	313.2	−305.8	717.8	802.5	415.9	−272.3	784.4
	(413.5)	(419.1)	(429.1)	(433.7)	(524.6)	(745.1)	(492.8)	(503.1)
82	285.1	175.4	−262.5	810.4	326.0	−244.3	−160.2	675.1
	(461.2)	(471.6)	(476.7)	(486.3)	(608.9)	(647.8)	(590.0)	(564.1)
83	96.0	419.5	−177.3	543.6	315.4	254.3	−53.6	462.3
	(512.6)	(538.1)	(531.5)	(523.2)	(720.0)	(767.5)	(643.4)	(638.9)
84	−76.8	−223.1	−123.3	641.3	−287.4	−718.6	−288.0	827.3
	(548.2)	(562.8)	(568.5)	(568.2)	(804.0)	(771.5)	(721.0)	(716.8)

Notes: Standard errors in parentheses.

The table shows the difference in earnings by lottery-determined draft-eligibility status. Eligibility ceilings are RSN 195 for men born in 1950, RSN 125 for men born in 1951, and RSN 95 for men born in 1952 and 1953.

Earnings data are from the Social Security Administration CWHS, described in the text and the Appendix.

self-selected on the basis of unobserved characteristics.

The draft lottery facilitates estimation of (1) because functions of randomly assigned lottery numbers provide instrumental variables that are correlated with s_i, but orthogonal to the error term, u_{it}. For example, one such instrument is a dummy variable, d_i, that equals one if the ith individual was draft eligible. Suppose that attention is restricted to a single cohort. Then, use of d and a constant as instrumental variables leads to the following estimator for α:

$$\hat{\alpha} = (\bar{y}^e - \bar{y}^n)/(\hat{p}^e - \hat{p}^n), \tag{2}$$

where $\hat{p}$ is the proportion of the cohort actually entering the military, $\bar{y}$ is mean earnings, and superscript e and superscript n denote the draft-eligible and draft-ineligible samples. Note that the numerator of (2) consists of estimates of the effect of draft eligibility plotted in Figure 2.

Intuitively, equation (2) simply adjusts earnings differences by draft-eligibility status for the fact that not all draft-eligible men actually served in the military, while some men who were not draft eligible voluntarily enlisted for service. The justification for estimation of the effects of military service in this manner is clear: it is assumed that nothing other than differences in the probability of being a veteran is responsible for differences in earnings by draft-eligibility status. This formula may also be recognized as an application of Abraham Wald's (1940) grouping method, where the data have been grouped by draft-eligibility status. Applications of this formula will therefore be referred to here as "Wald estimates."

In addition to draft-eligibility treatment effects, implementation of the Wald estimator requires estimates of $\hat{p}^e$ and $\hat{p}^n$. These estimates are tabulated from a special version of the 1984 Survey of Income and Program Participation (SIPP). The SIPP data used here were matched to an indicator of draft-eligibility status from information on birthdates included in the Census Bureau's in-house version of the SIPP file. Additional details on the SIPP data are provided in Section 7 of the Appendix.

In the upper panel of Table 2, the columns labeled $\hat{p}^e$ and $\hat{p}^n$ show probabilities of veteran status tabulated using the SIPP. Because of the small number of observations available for single-cohort statistics, each SIPP estimate is actually the average for three consecutive cohorts. For example, SIPP estimates assigned to men born in 1951 are based on data for men born in 1950, 1951, and 1952. The last column of Table 2, labeled $\hat{p}^e - \hat{p}^n$, shows the difference in the probability of military service by draft-eligibility status. Estimates of the effect of draft eligibility on veteran status for whites born 1950–52 range from 0.10 to 0.16. Thus, a rule of thumb for conversion of draft-eligibility treatment effects into estimates of the effects of military service is to multiply by $1/0.15 = 6\ 2/3$.

Wald estimates of the effect of military service for selected cohorts and years are presented in Table 3. The sample is restricted to the subset of whites born 1950–52 because the results in Table 2 suggest that this is the group for whom draft eligibility is most likely to be a useful instrument. Earnings variables are for 1981–84 because the impact of military service in these years represents a long-term effect. Furthermore, as a practical matter, both FICA and W-2 earnings data are likely to be more reliable in recent years—the FICA data because of increased employment coverage and the W-2 data because of improvements in data collection procedures.

Table 3 reports three sets of estimated draft-eligibility effects for use in the numerator of the Wald estimator. Column (1) presents estimates for FICA earnings and column (3) presents estimates for W-2 earnings; the figures in both of these columns are copied directly from Table 1. In addition, column (2) reports estimates for an earnings series constructed by applying a simple non-parametric correction for censoring to the FICA earnings data. The correction procedure is described in detail in Section 6 of the Appendix. Briefly, data are adjusted for censoring by using the fraction with recorded earnings at the taxable maximum, combined with mean earnings above the taxable maximum estimated from Current Population Surveys, to estimate population mean earnings from censored mean earnings. Note that the effects estimated using the adjusted data are usually bracketed by the effects estimated using the unadjusted FICA and W-2 data. Therefore, only Wald estimates constructed from the adjusted data are reported in the table.

Wald estimates for adjusted FICA earnings, reported in column (5) of Table 3, indicate that white veterans suffered an annual earnings loss of roughly \$2000 constant (1978) dollars of \$3,500 current dollars. This is approximately 15 percent of annual W-2

TABLE 2—VETERAN STATUS AND DRAFT ELIGIBILITY

Whites

Data Set	Cohort	Sample	P(Veteran)	$\hat{p}^e$	$\hat{p}^n$	$\hat{p}^e-\hat{p}^n$
SIPP (84)[a]	1950	351	0.2673	0.3527	0.1933	0.1594
			(0.0140)	(0.0325)	(0.0233)	(0.0400)
	1951	359	0.1973	0.2831	0.1468	0.1362
			(0.0127)	(0.0390)	(0.0180)	(0.0429)
	1952	336	0.1554	0.2310	0.1257	0.1053
			(0.0114)	(0.0473)	(0.0146)	(0.0495)
	1953	390	0.1298	0.1581	0.1153	0.0427
			(0.0106)	(0.0339)	(0.0152)	(0.0372)
DMDC/CWHS[b]	1950	16119	0.0633	0.0936	0.0279	0.0657
			(0.0019)	(0.0032)	(0.0019)	(0.0037)
	1951	16768	0.1176	0.2071	0.0708	0.1362
			(0.0025)	(0.0053)	(0.0024)	(0.0059)
	1952	17703	0.1515	0.2683	0.1102	0.1581
			(0.0027)	(0.0065)	(0.0027)	(0.0071)
	1953	17749	0.1343	0.1548	0.1268	0.0280
			(0.0026)	(0.0053)	(0.0029)	(0.0060)

Nonwhites

Data Set	Cohort	Sample	P(Veteran)	$\hat{p}^e$	$\hat{p}^n$	$\hat{p}^e-\hat{p}^n$
SIPP (84)[a]	1950	70	0.1625	0.1957	0.1354	0.0603
			(0.0292)	(0.0699)	(0.0491)	(0.0854)
	1951	63	0.1703	0.2014	0.1514	0.0500
			(0.0292)	(0.0827)	(0.0448)	(0.0940)
	1952	52	0.1332	0.1449	0.1287	0.0161
			(0.0275)	(0.1040)	(0.0373)	(0.1105)
	1953	55	0.1749	0.1980	0.1612	0.0367
			(0.0305)	(0.0865)	(0.0470)	(0.0984)
DMDC/CWHS[b]	1950	5447	0.0417	0.0548	0.0271	0.0276
			(0.0027)	(0.0042)	(0.0032)	(0.0053)
	1951	5258	0.0794	0.1173	0.0599	0.0574
			(0.0037)	(0.0076)	(0.0040)	(0.0086)
	1952	5493	0.0953	0.1439	0.0794	0.0644
			(0.0040)	(0.0095)	(0.0042)	(0.0104)
	1953	5303	0.0925	0.0984	0.0904	0.0080
			(0.0040)	(0.0079)	(0.0046)	(0.0092)

Notes: Standard errors in parentheses. $\hat{p}^e$ is the probability of being a veteran conditional on being draft eligible; $\hat{p}^n$ is the probability of being a veteran conditional on being ineligible.

[a] Wave I, Panel I of the 1984 Survey of Income and Program Participation. Probabilities are for service in the Vietnam era. Estimates are weighted by the SIPP sampling weight and smoothed over 3 cohorts.

[b] Defense Manpower Data Center Administrative Records' information on accessions, from 1970–73, combined with information on cohort size from the Social Security Administration Continuous Work History Sample.

compensation for white men between 1981 and 1984. The similarity of coefficient estimates across cohorts and years suggests that the Wald estimates provide a robust measure of the impact of military service. Taken individually, however, few of the estimates are statistically significant at conventional levels.[7]

[7] The asymptotic standard error of the Wald estimates is derived from the limiting distribution of

TABLE 3—WALD ESTIMATES

		Draft-Eligibility Effects in Current $				
		FICA Earnings	Adjusted FICA Earnings	Total W-2 Earnings	$\hat{p}^e - \hat{p}^n$	Service Effect in 1978 $
Cohort	Year	(1)	(2)	(3)	(4)	(5)
1950	1981	−435.8	−487.8	−589.6	0.159	−2,195.8
		(210.5)	(237.6)	(299.4)	(0.040)	(1,069.5)
	1982	−320.2	−396.1	−305.5		−1,678.3
		(235.8)	(281.7)	(345.4)		(1,193.6)
	1983	−349.5	−450.1	−512.9		−1,795.6
		(261.6)	(302.0)	(441.2)		(1,204.8)
	1984	−484.3	−638.7	−1,143.3		−2,517.7
		(286.8)	(336.5)	(492.2)		(1,326.5)
1951	1981	−358.3	−428.7	−71.6	0.136	−2,261.3
		(203.6)	(224.5)	(423.4)	(0.043)	(1,184.2)
	1982	−117.3	−278.5	−72.7		−1,386.6
		(229.1)	(264.1)	(372.1)		(1,312.1)
	1983	−314.0	−452.2	−896.5		-2,181.8
		(253.2)	(289.2)	(426.3)		(1,395.3)
	1984	−398.4	−573.3	−809.1		−2,647.9
		(279.2)	(331.1)	(380.9)		(1,529.2)
1952	1981	−342.8	−392.6	−440.5	0.105	−2,502.3
		(206.8)	(228.6)	(265.0)	(0.050)	(1,556.7)
	1982	−235.1	−255.2	−514.7		−1,626.5
		(232.3)	(264.5)	(296.5)		(1,685.8)
	1983	−437.7	−500.0	−915.7		−3,103.5
		(257.5)	(294.7)	(395.2)		(1,829.2)
	1984	−436.0	−560.0	−767.2		−3,323.8
		(281.9)	(330.1)	(376.0)		(1,959.3)

Notes: Standard errors in parentheses.
Columns (1) and (3) are taken from Table 1.
Column (2) reports draft-eligibility treatment effects on earnings adjusted for censoring at the FICA taxable maximum. The adjustment procedure is described in the Appendix. Column (4) reports SIPP estimates of the effect of draft eligibility on veteran status, taken from Table 2. Column (5) reports estimates of the effect of military service on civilian earnings is implied by columns (2) and (4).

B. *Efficient Instrumental Variables Estimates*

The Wald estimator is based solely on earnings differences by draft-eligibility status. A more efficient estimator exploits all the information on RSNs in the aggregate data by fitting earnings model (1) to observations on mean earnings for each group of five consecutive lottery numbers. Consider the following grouped version of (1), where $\bar{y}_{ctj}$ is mean earnings for members of cohort c at time t with lottery numbers in group j, and $\hat{p}_{cj}$ is the fraction of cohort c with lottery numbers in group j who served:

$$\bar{y}_{ctj} = \beta_c + \delta_t + \hat{p}_{cj}\alpha + \bar{u}_{ctj}. \tag{3}$$

Intuitively, estimation of (3) simply generalizes Wald's method to grouped data with more than two groups.

Generalized Least Squares (GLS) estimates of (3) may easily be shown to have an instrumental variables interpretation (Angrist, 1988). In this case, the instrument set includes dummy variables that indicate

$\sqrt{n}(\bar{y}^e - \bar{y}^n)/(\hat{p}^e - \hat{p}^n)$. The standard error is therefore equal to $1/(\hat{p}^e - \hat{p}^n)$ times the standard error of the numerator because the numerator has a nondegenerate limiting distribution, while $(\hat{p}^e - \hat{p}^n)$ converges to a constant. The same standard error formulas arise from application of conventional Instrumental Variables formulas.

groups of five consecutive lottery numbers for each race, cohort, and year of earnings. There are 73 dummy variables for a particular race, cohort, and year; the first indicates men with lottery numbers 1–5 and the 73rd indicates men with lottery numbers 360–365. Furthermore, the quadratic form minimized by the GLS estimator is an overidentification test statistic associated with the use of dummy variables as instruments. This statistic tests the exclusion of lottery number group dummies from equation (1). It may also be viewed as a measure of the goodness-of-fit of the cell means to equation (1).[8]

In principle, implementation of the estimation strategy based on (3) is straightforward—the estimates are simply coefficients from GLS regressions of mean Social Security earnings on estimates of $\hat{p}_{cj}$. The SIPP sample is too small to allow accurate estimation of a full set of $\hat{p}_{cj}$, however. Thus, a second set of probabilities was estimated from a combination of Defense Manpower Data Center (DMDC) administrative records and CWHS data on cohort size. Detailed descriptions of the DMDC administrative records may be found in Angrist (1989b). Briefly, the DMDC data show the total number of new entrants to the military by race, cohort, and lottery number from July 1970 through December 1973.

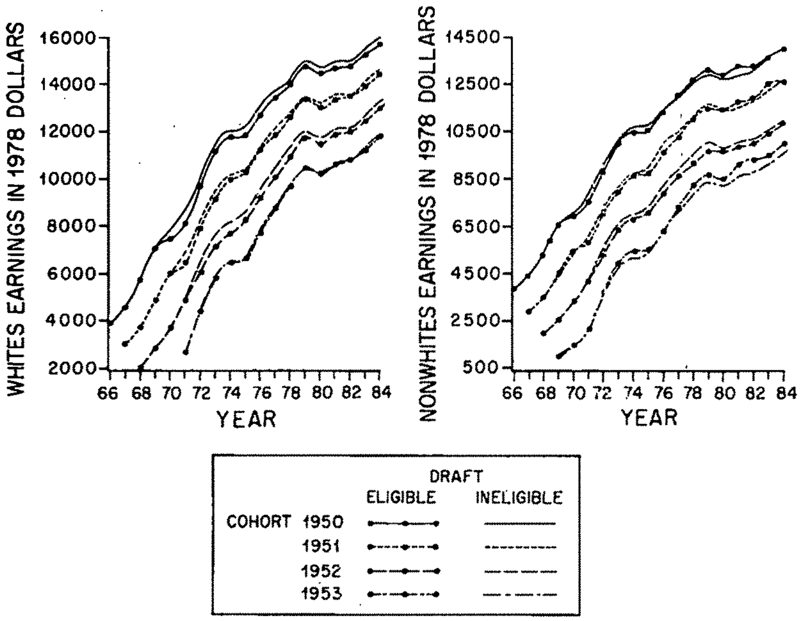
DMDC and CWHS administrative records are used to estimate $\hat{p}_{cj}$ by first counting the number of entrants to the military by race, cohort, and lottery number interval. These numbers are the numerator of the $\hat{p}_{cj}$. Estimates of overall cohort size, to be used in the denominator of $\hat{p}_{cj}$, are derived from the CWHS. Recall that the CWHS is a one percent sample, so that if the CWHS sampling frame is identified with the population at risk, an estimate of total cohort size is simply 100 times the CWHS cohort size. For example, to estimate the probability of being a veteran conditional on being draft eligible, the number of draft-eligible men in the DMDC data is divided by 100 times the number of men in the CWHS with lottery numbers below the induction ceiling. Standard errors for these estimates are computed by applying the usual formula for a binomial proportion.[9]

For comparison with the SIPP estimates, DMDC/CWHS estimates of $\hat{p}^e$ and $\hat{p}^n$ are reported in the lower panel of Table 2. These figures show that, with the exception of the 1950 cohort, the SIPP and DMDC/CWHS procedures give reasonably similar estimates of $\hat{p}^e - \hat{p}^n$. Inaccuracy of the DMDC/CWHS estimates for 1950 is a consequence of the fact that DMDC administrative records are unavailable before July 1970. Therefore, despite the limitations of the SIPP data, the SIPP must be used to construct probabilities for the 1950 cohort. The SIPP sample is too small to allow estimation of a complete set of $\hat{p}_{cj}$ for all lottery number cells in 1950. Consequently, SIPP estimates for 1950 are computed for only two cells, defined by draft-eligibility status, and CWHS earnings data for men born in 1950 are also grouped by draft eligibility. Thus, for each race and year, the sample used to estimate equation (3) includes 73 cell means for each of the three cohorts born from 1951–53, plus two cell means for the 1950 cohort.

A graphical version of equation (3) is depicted in Figure 3, which shows the relationship between probabilities of veteran status ($\hat{p}_{cj}$) and mean W-2 compensation in 1978 dollars ($\bar{y}_{ctj}$) between 1981 and 1984. Plotted in the figure are the average (over four years of earnings) residuals from a regression

[8] A general reference on overidentification testing is Whitney Newey (1985). See also Angrist (1988), where GLS on grouped data is shown to be the minimum variance linear combination of all the Wald estimators that can be computed from any division of grouped observations into linearly independent pairs. The overidentification test statistic for dummy variable instruments is also shown to be the same as the Wald statistic for equality of alternative Wald estimates.

[9] The formula used is $\sqrt{[\hat{p}(1-\hat{p})/n_c]}$, where $\hat{p}$ is the estimated proportion of servers and n_c is the number in the CWHS cohort. For example, 5749 draft eligible white men in the CWHS were born in 1951, and DMDC administrative records show that 119,062 draft-eligible white men born in 1951 served between July 1970 and December 1973. $\hat{p}^e$ is therefore 119,062/574,900 = 0.21, with estimated variance equal to (0.21 * 0.79)/5749.



Notes: The figure plots mean W-2 compensation in 1981–84 against probabilities of veteran status by cohort and groups of five consecutive lottery numbers for white men born 1950–53. Plotted points consist of the average residuals (over four years of earnings) from regressions on period and cohort effects. The slope of the least squares regression line drawn through the points is −2,384 with a standard error of 778, and is an estimate of α in the equation

$$\bar{y}_{ctj} = \beta_c + \delta_t + \hat{p}_{cj}\alpha + \bar{u}_{ctj}.$$

FIGURE 3. EARNINGS AND THE PROBABILITY OF VETERAN STATUS BY LOTTERY NUMBER

of earnings and probabilities on period and cohort effects.[10] Thus, the slope of the ordinary least squares regression line drawn through the points corresponds to an estimate of α. This slope is equal to −2,384 dollars, with a standard error of 778 dollars. An interesting feature of the figure is the apparent heteroscedasticity of the earnings residuals. Dispersion around the regression line is reduced for cells with high probabilities of veteran status. This heteroscedasticity also appears in comparisons (not shown here) of cell variances by draft-eligibility status; draft-eligible men have somewhat less variable earnings.

As pointed out earlier, estimation of (3) is the same as instrumental variables estimation of (1) using dummy variables as instruments. However, inference for the case where the estimation strategy is implemented by regressing CWHS mean earnings on DMDC/CWHS or SIPP probabilities is complicated by the use of data from multiple samples. Assuming that the samples used to calculate mean earnings and the sample used to calculate $\hat{p}_{cj}$ are independent, the optimal Two-Sample Instrumental Variables (TSIV)

[10] There are 221 points plotted in the figure: 4 years of earnings times 3 cohorts with 73 cells plus 4 years of earnings times one cohort with 2 cells (men born in 1950) = 884, divided by four to compute the average over years.

estimator has a simple form that may be briefly described as follows. Let $\bar{y}$ denote the vector of $\bar{y}_{ctj}$, $\hat{p}$ denote the vector of $\hat{p}_{cj}$, and $\bar{u}(\theta)$ denote the vector of $\bar{u}_{ctj}$, where θ in parentheses represents the dependence of residuals on the parameter vector. Also, let $V(\ )$ denote the covariance matrix of the argument. Then the optimal TSIV estimator chooses θ to minimize

$$\bar{u}(\theta)'\left[V(\bar{y})+\alpha^2 V(\hat{p})\right]^{-1}\bar{u}(\theta)\equiv m(\theta),$$

which is also the GLS minimand for (3).[11]

The minimized value of $m(\theta)$ is an overidentification test statistic for the validity of dummy variables as instruments. If some of these dummy variables are correlated with the regression error, then $m(\theta)$ should be large relative to a chi-square distribution with degrees of freedom equal to the difference between the number of instruments and the number of estimated parameters.

Table 4 presents two sets of TSIV estimates of equation (3) for 1981–84 earnings in 1978 dollars. Model 1 allows the effect of veteran status on earnings to vary by cohort, while Model 2 restricts estimated service effects to be the same across cohorts. Note that, as in Table 3, the heading "adjusted FICA earnings" refers to FICA-taxable earnings adjusted for censoring at the taxable maximum using the procedure described in section 6 of the Appendix.

The results in Table 4 show that white veterans born from 1950–52 suffered an annual earnings loss of between $1,500 and $2,100 constant (1978) dollars. These results generally are similar in magnitude to the Wald estimates reported in Table 3. Also, as in Table 3, regression estimates for adjusted FICA earnings tend to be bracketed by the results for unadjusted FICA and W-2 earnings. Although many of the estimates for individual cohorts in Model 1 are not significant, the combined estimates for whites in Model 2 are substantially larger than twice their standard errors. In contrast, results for nonwhites show no evidence of a statistically significant earnings loss to veterans.

The overidentification test statistics reported in Table 4 take on values less than their degrees of freedom, suggesting that the residuals in equation (1) are not correlated with lottery-based instruments.[12] It should be noted, however, that low values of the test statistics may indicate low power in a test with so many degrees of freedom. On the other hand, without a particular alternative hypothesis in mind, it seems natural to report the omnibus goodness-of-fit test.

Subtracting the test statistic for Model 1 from the test statistic for Model 2 gives a chi-square test for the restriction of equal treatment effects across cohorts. The set of restrictions imposed by equal treatment effects has three degrees of freedom. None of the chi-square statistics for Model 2 are larger than the corresponding statistics for Model 1 by as much as three, indicating that

[11] See Angrist (1989d) for details. Estimates of α for use in the formula for Φ were computed by weighted least squares using the inverse of the sampling variance of the $\bar{y}_{ctj}$ as weights. Estimates of $V(\bar{y})$ and $V(\hat{p})$ are discussed in Section 5 of the Appendix. Note that the TSIV estimator may also be motivated as an application of Optimal Minimum Distance (OMD) techniques such as those described by Gary Chamberlain (1982). Ignoring period and cohort effects, OMD estimates for the current problem are tabulated by choosing α and p to minimize

$$q(\alpha,p)=$$

$$\begin{bmatrix}\bar{y}-p\alpha\\ \hat{p}-p\end{bmatrix}'\begin{bmatrix}V(\bar{y}) & 0\\ 0 & V(\hat{p})\end{bmatrix}^{-1}\begin{bmatrix}\bar{y}-p\alpha\\ \hat{p}-p\end{bmatrix}.$$

By concentrating out the estimate of p, it is easy to show that $q(\alpha,p)=m(\theta)$.

[12] Degrees of freedom for the overidentification tests are calculated as follows. For each race, the data consist of four years of earnings for three cohorts with 73 lottery number cells each. The fourth cohort, men born in 1950, has four years of earnings with 2 lottery number cells each. This gives a total of 884 cells or, equivalently, 884 categorical instruments. Model 1 includes 4 cohort dummies, 3 year dummies, and 4 treatment effects. 884 minus 11 parameters gives 873 degrees of freedom. Model 2 has 3 fewer parameters than model 1 and consequently the chi-square statistic for model 2 has 876 degrees of freedom.

TABLE 4—TWO-STAGE INSTRUMENTAL VARIABLES ESTIMATES

Cohort	FICA Taxable Earnings	Adjusted FICA Earnings	Total W-2 Compensation
Whites			
Model 1			
1950	−1709.2	−2093.7	−1895.0
	(946.8)	(1108.8)	(1333.1)
1951	−1457.1	−1983.7	−2431.4
	(959.3)	(1036.1)	(1152.1)
1952	−1724.0	−1943.0	−2058.7
	(863.1)	(927.2)	(1001.9)
1953	1223.8	900.7	−488.6
	(3232.1)	(3505.3)	(3936.0)
$\chi^2(873)$	578.3	630.3	569.5
Model 2			
1950–53	−1562.9	−1920.4	−2094.5
	(521.8)	(575.9)	(646.3)
$\chi^2(876)$	579.1	631.0	569.7
Nonwhites			
Model 1			
1950	3893.7	3891.9	5711.8
	(5358.5)	(6244.5)	(7206.0)
1951	−891.3	−333.4	2609.0
	(4397.1)	(4664.2)	(4894.6)
1952	−3182.9	−3457.7	−3068.0
	(3997.4)	(4195.2)	(4229.2)
1953	−5928.3	−8571.4	−6325.8
	(10296.3)	(10697.1)	(11410.6)
$\chi^2(873)$	616.7	681.7	693.6
Model 2			
1950–53	−643.3	−999.7	366.7
	(2407.5)	(2602.5)	(2734.2)
$\chi^2(876)$	618.4	683.4	695.6

Notes: Standard errors in parentheses.
The table shows estimates of the effect of military service on average 1981–84 earnings in 1978 dollars. The estimation method is optimally weighted Two-Sample Instrumental Variables, described in the text. FICA and W-2 earnings are from the Social Security CWHS. The adjusted FICA series is described in the Appendix.

the estimated treatment effects are not statistically different across cohorts.

IV. Military Service and Loss of Labor Market Experience

The simplest explanation for a veteran earnings penalty is that military experience is a poor substitute for lost civilian labor market experience. As evidence for this hypothesis, Griliches and Mason (1972) show that the longer they were in the military, the less veterans earn relative to nonveterans. A test of the loss-of-experience hypothesis is developed here using the functional form commonly employed in empirical studies of human capital. The earnings function motivated by the theory of human capital is loglinear in years of schooling and log-quadratic in years of labor market experience. This functional form puts testable restrictions on the time-series of earnings differences by veteran status.[13]

Adapting the human capital earnings function for the problem at hand, the earnings of individual i in cohort c at time t may be written

$$\text{(4a)} \qquad y_{cti} = \delta_t + w_i\delta_0 + \beta_0 x_{ict} + \gamma x_{ict}^2 + u_{it},$$

where y_{cti} now denotes log earnings, δ_t is a time-varying intercept, β_0, γ, and δ_0 are parameters. x_{ict} is the civilian labor market experience of man i in cohort c at time t, taken here to be equal to $[t-(c+18)-w_i-s_i l]$, where w_i is the deviation of i's schooling from the sample mean level of schooling, and l is years of military experience for veterans. As before, s_i is a dummy variable that indicates military service.

To focus on parameters that can be estimated using Social Security data, equation (4) is rewritten as

$$\text{(4b)} \qquad y_{cti} = \delta_{it} + \beta_0(x_{ct} - s_i l) + \gamma(x_{ct} - s_i l)^2 - (2\gamma w_i x_{ct} - 2\gamma l w_i s_i) + u_{it},$$

where $x_{ct} = t-(c+18)$ and $\delta_{it} = \delta_t + w_i(\delta_0 - \beta_0) + \gamma w_i^2$. Now, as in the previous analysis, assume that schooling does not vary by lottery number. Assume also that schooling is independent of cohort—this seems reasonable for the small cohort range considered here. Finally, to focus solely on the loss of labor market experience, assume that schooling is independent of veteran status. Then using dummy instrumental variables to group equation (4b) by cohort, year, and lottery number, average log earnings for members of cohort c at time t in lottery-number cell j are

$$\text{(5)} \qquad \bar y_{ctj} = \delta_t + \beta_0 x_{ct} + \gamma x_{ct}^2 - [\beta_0 l - \gamma l^2]\,\hat p_{cj} - [2\gamma l](\hat p_{cj} x_{ct}) + \bar u_{ct},$$

where δ_t now includes the period mean of δ_{it}.[14]

A generalization of model (5) allows the linear experience term to vary with veteran status by letting the slope for individual i be $\beta_i = \beta_0 + \beta_1 s_i$. In this case, mean cell earnings are characterized by

$$\text{(6)} \qquad \bar y_{ctj} = \delta_t + \beta_0 x_{ct} + \gamma x_{ct}^2 - [\beta_0 l - \gamma l^2 + \beta_1 l]\,\hat p_{cj} - [2\gamma l - \beta_1](\hat p_{cj} x_{ct}) + \bar u_{ct}.$$

Models (5) and (6) both have the following reduced form in terms of unrestricted regression coefficients:

$$\text{(7)} \qquad \bar y_{ctj} = \delta_t + \beta_0 x_{ct} + \gamma x_{ct}^2 + \pi_1 \hat p_{cj} + \pi_2(\hat p_{cj}^* x_{ct}) + \bar u_{ct}.$$

Note that the reduced form veteran effect is $\alpha_{ct} \equiv \pi_1 + \pi_2 x_{ct}$. Thus, these models parameterize a time-varying veteran status coefficient as a linear function of labor market

[13] See Mincer (1974) for theoretical justification of the human capital earnings function. A recent survey of the human capital literature is Willis (1986).

[14] Averaging over c, t, and j eliminates $(\partial\gamma w_i x_{ct} - \partial\gamma w_i s_i)$ because w_i is orthogonal to x_{ct} and s_i by assumption. Using the fact that $E(s_i|c,j) = E(s_i^2|c,j) = p_{cj}$, (4) simplifies to (5). Note that (5), (6), and (7) are not estimable if allowance need be made for cohort as well as period effects. Qualitatively similar estimates to those reported below were obtained when δ_t was dropped in favor of cohort effects, although the goodness-of-fit test leads to rejection of models without period effects.

TABLE 5—EARNINGS FUNCTION MODELS FOR THE VETERAN EFFECT, WHITES BORN 1950–52

Parameter	Model (5): Loss of Experience (1)	Model (6): Loss of Experience, Reduced Growth Rate (2)	Model (7): Unrestricted Reduced Form (3)
Experience Slope, β_0	0.1022	0.1016	0.1016
	(0.007)	(0.007)	(0.007)
Experience Squared, γ	−0.0027	−0.0025	−0.0025
	(0.0003)	(0.0003)	(0.0003)
Veteran Effect on Slope, β_1		−0.0035	
		(0.0023)	
Veteran Loss of Experience, l	2.08	1.84	
	(0.38)	(0.43)	
$\pi_1 = -[\beta_0 l - \gamma l^2 + \beta_1 l]$			−0.189
			(0.052)
$\pi_2 = -[2\gamma l - \beta_1]$			0.006
			(0.004)
Age at Which Reduced Form			50.1
Veteran Effect $(\pi_1 + \pi_2 x_{ct}) = 0$			(15.9)
χ^2(dof)	1.41(1)		813.57(1247)

Notes: Standard errors in parentheses.

The table reports estimates of experience-earnings profiles that include parameters for the effect of veteran status. Estimates are of equations (5), (6) and (7) in the text. The estimating sample includes FICA taxable earnings from 1975–84 for men born 1950, 1976–84 earnings for men born in 1951, and 1977–84 earnings for men born 1952. The estimation method is optimally weighted Two-Sample Instrumental Variables for a nonlinear model in columns (1) and (2), and for a linear model in column (3).

experience. Excluding the time-varying intercept, model (6) contains four structural parameters; β_0, β_1, γ, and l, and four reduced form parameters; β_0, γ, π_1 and π_2. Model (5) imposes one testable restriction on the reduced form by setting $\beta_1 = 0$.[15]

Table 5 shows results from nonlinear GLS estimation of (5) and (6), and results from Linear GLS estimation of (7), using data on the real FICA earnings of white men born from 1950 to 1952. The weighting matrix used in estimation was derived in a manner similar to the weighting matrix used to construct the estimates in Table 4.[16] Because earnings functions are commonly fit in logs, the dependent variable is taken to be the log of mean earnings for each cell. The log of the mean is not the same as the mean of the log, but the CWHS data set does not contain the mean of log earnings. If earnings are approximately lognormally distributed, use of the log of the mean will provide a reasonable approximation. In practice, estimates of

[15] A third model is derived by letting both γ and β vary with veteran status. This model leads to a reduced form similar to (7), with the only modification being the addition of a linear term of the form $\pi_3(\hat{p}_{cj} * x_{ct}^2)$. In the empirical work, however, no evidence was found that such a term belongs in the earnings function reduced form.

[16] The only modification arises from the fact that for equations (5)–(7), reduced form treatment effects appearing in the weighting matrix are time varying. Let Π_c denote the vector of α_{ct} for the time-series of earnings by cohort c and suppose that each time-series is of length T. Then the second term in the optimal weighting matrix has the following block corresponding to the time series of earnings for lottery number cell j of cohort c:

$$(1/n_{cj})\Pi_c \times \left\{ [e_T e_{T'}] \otimes \left[\hat{p}_{cj}(1-\hat{p}_{cj}) \right] \right\} \Pi_c',$$

where e_T is a vector of T 1's. In practice, Π_c is replaced by weighted least squares estimates (weights are the inverse sampling variances of $\bar{y}_{ctj}$) of the reduced form equation, (7). Estimates of models for the log of earnings replace $V(\bar{y})$ with delta-method estimates of the covariance matrix of log($\bar{y}$).

models in levels resulted in inferences similar to those arising from estimates of models in logs.

The sample used to estimate equations (5)–(7) begins in the fifth year after the lottery in which members of the cohort were drafted. For example, the sample begins in 1975 for men born in 1950. This allows for three years of service and one year of readjustment to civilian life. Median length of service of Vietnam era veterans was 37 months (Veterans Administration 1981b, p. 16). Evidence from Table 4 suggests that veteran effects are essentially zero for the 1953 cohort, so it was excluded from the estimation.

Estimates of models (5), (6), and (7) are presented in columns (1), (2), and (3) of Table 5. The chi-square statistic in column (3) is an overidentification test statistic for the overall goodness-of-fit of the reduced form and the chi-square statistic in column (1) is for the restriction $\beta_1 = 0$. Restricting β_1 to be zero does not affect the overall fit. The effect of veteran status on earnings growth, although negative, is not statistically significant. The sum of the test statistics in columns (1) and (3) is an overidentification test statistic for the overall goodness-of-fit of model (5). This statistic also takes on a value less than its degrees of freedom, indicating that the simplest loss-of-experience model is not at odds with the data.

The loss of experience estimated using model (5) is approximately 2 years. This is somewhat low relative to the median length of service, suggesting that military experience may be a partial substitute for civilian experience. Table 5 also shows the age at which the reduced form veteran effect finally reaches zero. This occurs when $x_{ct} = -\pi_1/\pi_2$, so that $\alpha_{ct} = 0$. The reduced form estimates in column (3) imply that the loss of earnings to veterans decays to zero around age 50 with a standard error[17] of 16.

V. Caveats

The consistency of lottery-based estimates of the effects of military service turns on two key assumptions. First, earnings model (1) must be an accurate representation of the impact of military service. Second, functions of the draft lottery must be valid instruments for s_i in the earnings regressions. Three models leading to the failure of these assumptions are discussed briefly below. The first two, incorporating treatment effect heterogeneity and missing covariates, merely result in a reinterpretation of the estimates. The third model, allowing for earnings-modifying draft-avoidance behavior, calls into question the assumption that earnings differences by lottery number can be attributed solely to veteran status.

A. *Treatment Effect Heterogeneity*

The effect of veteran status can only be estimated for those who served in the armed forces. If veterans are more or less likely to benefit from military service than the rest of the population, then the estimated veteran status coefficient does not characterize the impact of veteran status on a random sample. This problem is formalized in the random coefficients model for treatment effect heterogeneity:

$$y_{cti} = \beta_c + \delta_t + s_i\alpha_i + u_{it}$$

$$\alpha_i = \alpha_0 + \varepsilon_i,$$

where α_i is the effect of military service on the earnings of person i, with population mean equal to α_0. If ε_i is uncorrelated with the instruments, an instrumental variables estimator of the random coefficients model identifies $\alpha_0 + E(\varepsilon_i|s_i = 1)$, and not α_0 (James Heckman and Richard Robb, 1985).

Related to the problem of treatment effect heterogeneity is the fact that not all Vietnam era accessions to the military were induced by the draft. A substantial fraction of enlistments were made by "true volunteers" who would have volunteered for service in the absence of a draft (Tarr, 1981). Suppose that the impact of military service on the earnings of true volunteers differs from the im-

[17] Delta method standard errors are given by the square root of $(\sigma_1/\pi_2^2) - (2\sigma_{12}\pi_1/\pi_2^3) + (\sigma_2\pi_1^2/\pi_2^4)$, where σ_1, σ_2, and σ_{12} are the elements of the covariance matrix of the estimated π_1 and π_2. σ_{12} is estimated to be -0.00019, and the square roots of σ_1 and σ_2 appear in Table 5.

pact on the earnings of draftees and men who enlisted because of the draft. Then using functions of the draft lottery as instruments will only identify the effect of military service for the latter group. To see this, let f_i be an indicator for the veteran status of draftees and draft-motivated volunteers and let g_i be an indicator of veteran status for true volunteers. Write

$$\alpha_i s_i = \alpha_f f_i + \alpha_g g_i$$

for the treatment effect experienced by i. Note that functions of draft lottery numbers will only be correlated with f_i. Consequently, α_f is identified by lottery-based IV estimators, but $\alpha_g g_i$ becomes part of the regression error.

B. *The Absence of Covariates*

An additional problem arises because Social Security data contain no information on covariates other than race and age. This may be of concern if the impact of veteran status is primarily through its effect on covariates. For example, veterans might have a higher level of educational attainment because of financial aid available through the GI Bill. In the absence of data on education, estimated veteran effects confound the "pure" effect of military service with its effect on education. Formally, the need for covariates may be represented by replacing δ_t with $x_i\delta$ in (1). In this case, instrumental variables estimates identify

$$\bar{\alpha} = \left[E(x_i|s_i=1) - E(x_i|s_i=0)\right]\delta + \alpha.$$

In many applications, however, it may be $\bar{\alpha}$ that is actually the parameter of interest. For example, the fact that veteran status influences civilian earnings primarily through its influence on third variables might be of little importance for issues related to veterans compensation.

C. *Earnings-Modifying Draft Avoidance Behavior*

Perhaps the most serious problem arises if the risk of induction affected earnings for some other reason than through an effect on the probability of military service. For example, it is sometimes argued that during the Vietnam era, students went to college to avoid the draft and that educational standards were reduced to avoid having to flunk students out of school. Lawrence Baskir and William Strauss (1978) claim that Vietnam era college enrollment was 6–7 percent higher than normal because of the draft.

If draft-avoidance behavior is correlated with lottery numbers and with variables related to earnings besides veteran status, then lottery-based instruments will be correlated with the regression error in equation (1). Such correlation will bias estimates of the effects of military service constructed using the lottery. But in a previous study using micro data (Angrist, 1989a), specification tests provided no evidence of a relationship between lottery numbers and characteristics other than veteran status. The overidentification test statistics reported here also show no evidence of omitted variables bias. Finally, even if having a low lottery number is correlated with a tendency to stay in school, the fact that earnings rise with schooling implies that lottery-based estimates of the effect of veteran status will tend to underestimate the true effect.

VI. Conclusions

Estimates based on the draft lottery indicate that as much as ten years after their discharge from service, white veterans who served at the close of the Vietnam era earned substantially less than nonveterans. The annual earnings loss to white veterans is on the order of $3,500 current dollars, or roughly 15 percent of yearly wage and salary earnings in the early 1980s. In contrast, the estimated veteran effects for nonwhites are not statistically significant.

In light of the results reported here, some more conventional estimates of the effect of Vietnam era veteran status do not appear to be too far off the mark. Rosen and Taubman (1982) report estimates close to these, finding a 19 percent annual earnings loss to Vietnam era veterans. Crane and Wise (1987) find an 11 percent reduction in 1979 weekly earnings. On the other hand, Mark Berger and Barry Hirsch (1983) find essentially no effect on 1977 weekly earnings, and Angrist (1989a) reports OLS coefficients of zero when using the National Longitudinal Survey (NLS) to estimate the effects of veteran status on hourly wages. In contrast to the OLS estimates, lottery-based estimates from the NLS indicate that white veterans had lower hourly wages in 1981 than their nonveteran counterparts. Similar results from the SIPP are reported in Angrist (1989c). Thus, lottery-based estimates from a variety of sources provide conclusive evidence that white Vietnam veterans were disadvantaged by their service.[18]

This paper also proposes a simple explanation for the loss of earnings to white veterans: they earn less because their military experience is only a partial substitute for the civilian labor market experience lost while in the armed forces. Goodness-of-fit tests suggest that for whites, the time-series of veteran status coefficients is consistent with this hypothesis. Experience-earnings profiles estimated using Social Security data imply that white veterans suffered an earnings reduction equivalent to the loss of two years civilian labor market experience.

The analysis reported here leads naturally to further research on a number of topics. One of these is the question of alternatives to the loss-of-experience explanation for the reduction in white veterans' earnings. Veteran status may be a screening device, as suggested by DeTray (1982), or there may be cohort size effects such as those discussed by Finis Welch (1979). Because the Social Security data include information on variances, testable implications of these theories might also include restrictions on second, as well as first, moments. Another question for future research is whether draft eligibility affected educational and career plans independently of its effect on military service. The lottery may provide a useful tool for research on changing educational attainment in the 1960s and 1970s.

Finally, there remains the question of reconciling the loss of earnings to Vietnam era veterans with the apparent benefits of military service to veterans of World War II and other eras (Rosen and Taubman, 1982; Berger and Hirsch, 1983). Elsewhere, Alan Krueger and I have argued that the need for reconciliation is, at least in part, illusory (Angrist and Krueger, 1989). Although OLS regressions usually show that the effect of World War II veteran status is large, positive, and significant, these results may actually be a consequence of selection bias. By exploiting the fact that World War II veteran status is also correlated with exact date of birth, we have implemented an instrumental variables estimation strategy similar in spirit to the one used here. The results of this procedure indicate that the true impact of World War II veteran status on earnings is no larger than zero and may well be negative.

APPENDIX: DATA SOURCES AND METHODS

1. CWHS Data Collection[19]

The Social Security Administration maintains the earnings histories of covered employees in a data base known as the Summary Earnings Record (SER). Approximately one year after the SER has been updated with the latest year's earnings, a one percent sample of earnings histories is extracted. The sampling frame consists of all issued Social Security numbers, and the sample is stratified using the regional information coded in the numbers.

Prior to 1978, the FICA taxable earnings of employees were reported to the SSA by employers on a quarterly basis. Self-employed workers report their earnings annually on schedule SE of Internal Revenue Service (IRS) Form 1040, which is forwarded to the SSA by the IRS. Since 1978, employers have no longer been required to make quarterly reports. Instead they file IRS form W-2 with the SSA, as well as with the IRS, on an annual basis.

After 1978, all earnings, including those above the FICA taxable maximum, are to be reported to the SSA on form W-2. Furthermore, all employers are required

[18] Lottery-based estimates from the SIPP and NLS suggest that nonwhite veterans may have higher monthly and hourly earnings than comparable nonveterans. The estimates for nonwhites are too imprecise, however, to be viewed as conclusive.

[19] This section draws on Warren Buckler and Cresston Smith (1984), U.S. Department of Health and Human Services (1987), and personal correspondence and conversations with Buckler and Smith.

to file W-2s with the SSA regardless of whether their employees are engaged in FICA taxable employment. In practice, however, many employers do not report the earnings of those engaged in non-FICA-taxable employment. A further shortcoming of the W-2 series is the poor quality of the data during the first years of annual reporting.

2. Coverage and Truncation of the CWHS Earnings Series

FICA coverage includes most wage and salary and self-employment earnings. For the sample period used here, the most important coverage exceptions are the majority of federal civilian employees, some state and local government employees, some agricultural and domestic workers, and the employees of some nonprofit organizations. A view of coverage by industry in 1981 is given in Robert Meyer (1985, Table 2.1). Meyer's figures for state, local, and civilian federal government employees imply that by 1981, roughly 58 percent of all civilian government workers were covered.

FICA taxable maximums are reported in Appendix Table A1. The combined effects of limited coverage and censoring at the taxable maximum are conveniently summarized by the percentage of all earnings that are reported to the SSA. These statistics, reported in Department of Health and Human Services (1987, Table 30), show that after 1981 over 90 percent of wage and salary earnings and over 75 percent of self-employment earnings were reported to the SSA.

The W-2 earnings series excludes earnings from self-employment. Unpublished estimates indicate that roughly 6.4 percent of men born between 1944 and 1953 with nonzero earnings in 1984 had self-employment earnings only (figures from private correspondence with SSA employees). Other differences between FICA and W-2 earnings coverage are described by Mary Millea and Beth Kilss (1980).

3. Matching Dates of Birth to the CWHS

In the CWHS, information on race and sex is obtained from a computerized record of applications for a Social Security number called the NUMIDENT file. For this project, dates of birth were also matched from the NUMIDENT file to the CWHS. Draft lottery numbers were then matched to the dates of birth using lottery number tables in Selective Service System Semiannual Reports for 1969–73. A small number of individuals were discarded from the final data set because there was no information on either their sex, race, or exact date of birth.

4. CWHS Descriptive Statistics

Descriptive statistics for the CWHS cohorts studied here are reported in Table A1. The combined statistics for four cohorts were constructed by computing weighted averages of cohort means. Unless otherwise noted, statistics in the table refer to men with positive earnings. Sample sizes decline over time due to attrition from mortality.

The descriptive statistics indicate that after 1972 the fraction of men in the sample with zero FICA earnings varies roughly between 15 and 22 percent for whites and between 33 and 36 percent for blacks. To evaluate these figures, note that the author's tabulations show that roughly 10 percent of white men in these cohorts have zero recorded wage and salary earnings in the late 1970s' Current Population Surveys. Suppose that of the 90 percent who work, 12 percent are in the uncovered sector so that only 79 percent of the cohort may be expected to have positive FICA wage and salary earnings. Adding an estimated 5 percent who only have FICA self-employment earnings implies that 16 percent should have zero FICA earnings of any type. Thus, 14–18 percent of CWHS earnings being zero for whites between 1973 and 1980 does not seem unreasonable.

In recent years, the fraction with zero earnings appears to be too high to be accounted for by labor force participation or employment in the uncovered sector. This is probably because of the long delay in filing and recording Social Security taxable earnings. In their analysis of Social Security data, Card and Sullivan (1988) also note the problems caused by filing delay. Problems of undercoverage and filing delay may be especially severe for nonwhites. The 15 percent of CPS nonwhites with zero wage and salary earnings is not large enough to explain the approximately 34 percent of nonwhites with zero FICA earnings in the CWHS.

The fraction with FICA earnings at the taxable maximum is more variable than the fraction with zero earnings, ranging from 3 to 15 percent for whites and between 2 to 10 percent for blacks. The FICA earnings of men with multiple employers can exceed the taxable maximum because reported earnings are censored by source. The fraction of men with FICA earnings above the taxable maximum is around 1–2 percent. The W-2 and FICA earnings series show roughly equal fractions at or above the FICA taxable maximum, suggesting that both variables are drawn from the same underlying distribution. But problems with early years of the W-2 series are clearly reflected in the sample statistics. For example, the standard deviation of whites' W-2 earnings in 1978 is six times the mean and does not fall below the mean until 1981. Another disturbing feature of the W-2 series is that *nominal* earnings appear to fall from 1978–80. The W-2 series also has a substantially higher fraction of zeros than the FICA series does. However, part of this difference is caused by the inclusion of self-employment earnings in the FICA series. Also, generally there are some individuals with FICA taxable earnings but no federally taxable compensation (Millea and Kilss, 1980).

5. Covariance Estimates for Social Security Earnings and for $\hat{p}_{cj}$

Information on second moments in the aggregated CWHS is restricted to variances. Therefore, off-diagonal elements of the covariance matrix of Social Security earnings must be estimated. Recall that $\bar{y}$ denotes the vector of $\bar{y}_{ctj}$, where c indexes cohort, t indexes year of earnings, and j indexes lottery number cells. The covariance matrix of $\bar{y}$ is block diagonal, with nonzero elements for the covariance between $\bar{y}_{ctj}$ and $\bar{y}_{ckj}$, and zeros everywhere else. Thus, there is only correlation between elements of the time-series of earnings for a particular cohort and lottery number group. The in-

TABLE A1—DESCRIPTIVE STATISTICS FOR MEN BORN 1950–53

Race	Year	N	Taxable Maximum	FICA Earnings	FICA at Limit[a,b]	FICA Zeros	W-2 Earnings[c]	W-2 at Limit[a]	W-2 Zeros[a]
White	69	68,407	7,800	1,473 (1,457)	0.003	0.321			
	70	68,339	7,800	1,977 (1,825)	0.010	0.241			
	71	68,244	7,800	2,581 (2,233)	0.030	0.194			
	72	68,154	9,000	3,614 (2,802)	0.040	0.157			
	73	68,053	10,800	4,738 (3,460)	0.046	0.141			
	74	67,966	13,200	5,727 (4,170)	0.034	0.145			
	75	67,882	14,100	6,459 (4,843)	0.053	0.169			
	76	67,794	15,300	7,698 (5,548)	0.078	0.167			
	77	67,691	16,500	8,974 (6,206)	0.107	0.163			
	78	67,598	17,700	10,441 (7,050)	0.149	0.165	15,435 (91,032)	0.167	0.244
	79	67,503	22,900	12,388 (8,455)	0.092	0.166	14,786 (65,359)	0.097	0.261
	80	67,413	25,900	13,769 (9,678)	0.092	0.175	14,561 (28,180)	0.096	0.272
	81	67,316	29,700	15,641 (11,129)	0.089	0.183	16,363 (15,295)	0.092	0.272
	82	67,265	32,400	16,743 (12,371)	0.092	0.200	17,907 (16,298)	0.093	0.282
	83	67,190	35,700	18,046 (13,621)	0.089	0.210	19,595 (22,470)	0.089	0.275
	84	67,114	37,800	19,717 (14,883)	0.103	0.219	21,595 (20,856)	0.101	0.281
Nonwhite	69	21,514	7,800	1,306 (1,471)	0.002	0.464			
	70	21,501	7,800	1,711 (1,865)	0.007	0.417			
	71	21,475	7,800	2,197 (2,275)	0.020	0.388			
	72	21,443	9,000	3,072 (2,922)	0.025	0.352			
	73	21,412	10,800	3,992 (3,609)	0.029	0.330			
	74	21,380	13,200	4,802 (4,359)	0.021	0.336			
	75	21,339	14,100	5,404 (5,015)	0.031	0.359			
	76	21,310	15,300	6,453 (5,836)	0.049	0.350			
	77	21,275	16,500	7,510 (6,616)	0.071	0.341			
	78	21,243	17,700	8,751 (7,541)	0.096	0.338	13,439 (63,722)	0.116	0.391
	79	21,200	22,900	10,262 (8,938)	0.061	0.329	13,581 (63,244)	0.067	0.393
	80	21,167	25,900	11,405 (10,147)	0.063	0.331	11,716 (13,895)	0.067	0.397
	81	21,130	29,700	12,986 (11,628)	0.063	0.335	13,421 (14,644)	0.063	0.388
	82	21,109	32,400	14,045 (12,849)	0.067	0.354	14,983 (16,112)	0.065	0.408
	83	21,077	35,700	15,101 (14,167)	0.067	0.358	16,271 (19,089)	0.065	0.396
	84	21,042	37,800	16,391 (15,237)	0.077	0.353	17,905 (20,784)	0.073	0.387

Notes: Statistics are from the Social Security Administration CWHS. Standard deviations of earnings in parentheses. Amounts are in current dollars. Sample statistics are weighted averages of cells for each race, year of birth, and five consecutive lottery numbers.

[a] Fractions at limit are fractions of nonzero earnings at or above FICA taxable maximum. Fractions zero are fractions of all nondecedents with zero earnings.

[b] FICA earnings are wage and salary and self-employment earnings in Social Security taxable employment. FICA taxable earnings are censored at the taxable maximum except for those with multiple sources. Multiple sources are censored by source.

[c] W-2 earnings are total W-2 form wage and salary compensation, not censored at the Social Security taxable maximum. W-2 earnings do not include earnings from self-employment.

tertemporal correlation structure is assumed constant across lottery number groups so that the 73 cells available for each race and cohort may be used to estimate correlation matrices for all race-cohort combinations. Correlations are converted to covariances using the within cell variances available from the CWHS data. This procedure is also applied to the adjusted FICA series described in Section 6 of the Appendix, with the modification that adjusted standard errors (diagonal elements of Ω) are used to convert correlations to covariances.

The covariance matrix of $\hat{p}$ is also block diagonal, with elements equal to the variance of $\hat{p}_{cj}$ in every element of the block corresponding to the time series for cohort c and lottery cell j. The variance of $\hat{p}_{cj}$ is estimated using the standard formula for an estimated proportion. The sample size in the formula is taken to be the size of the SIPP cohort for those born in 1950 and the size of the CWHS cohort for those born from 1951 to 53.

6. FICA Earnings Adjusted for Censoring

For economy of notation, in this section all cells are indexed by j. The relationship between the expectation of censored earnings and the expectation of uncensored earnings for cell j is given by

$$\mu_j^0 = \mu_j^c + p_j^l\left(\mu_j^l - L_j\right),$$

where μ_j^0 is the mean of uncensored earnings, μ_j^c is the mean of censored earnings, μ_j^l is the mean of earnings above the taxable maximum, L_j is the taxable maximum, and p_j^l is the fraction with earnings at or above the taxable maximum.[20]

The FICA earnings series is adjusted for censoring by applying this formula using estimates of μ_j^l tabulated from March Current Population Surveys (CPS) for each year, race, and cohort. Although this procedure involves no parametric distributional assumptions, the adjustment is only approximate because the CPS estimates of μ_j^l do not vary by lottery number. However, the adjustment does incorporate variation in p_j^l by lottery number.

Data on earnings above the FICA taxable maximum (μ_j^l) are taken from the Mare-Winship March CPS Uniform files. Cohort was determined on the basis of age in 1985; ages 34–35 were assigned 1950, ages 33–34 were assigned 1951, ages 32–33 were assigned 1952, and ages 31–32 were assigned 1953. The CPS reports wage and salary earnings in the year preceding the survey year. To compute the adjusted 1981–84 earnings series used in Tables 3 and 4, data from the CPSs for 1982–85 were used to compute separate a μ_j^l by cohort and race for each year. Additional details and summary statistics for the CPS data are reported in Appendix A of Angrist (1989c). It should be noted that CPS earnings data are also censored—at \$75,000 for 1981–83 earnings and at \$100,000 for 1984 earnings.

Standard errors for the adjusted series are calculated as follows. Let $m = [m_1' \, m_2' \, m_3']'$ denote the vector of sample moments corresponding to μ^c, p^l, and μ^l and let $\Sigma_{ij} (i, j = 1,2,3)$ denote the corresponding blocks of the covariance matrix of m. The covariance matrix of m is assumed to be given by

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} & 0 \\ \Sigma_{12} & \Sigma_{22} & 0 \\ 0 & 0 & \Sigma_{33} \end{pmatrix}.$$

The delta-method covariance matrix for the vector of adjusted earnings, $m_1 + m_2(m_3 - L)$, is

$$\Omega = \Sigma_{11} + 2\Sigma_{12}(m_3 - L) + \Sigma_{22}(m_3 - L)^2 + m_2^2\Sigma_{33}.$$

Only the diagonal elements of Ω are estimated using the above formula. Estimates of diagonal elements of Σ_{11} and Σ_{22} are available from the CWHS cell statistics, while diagonal elements of Σ_{33} are estimated from the CPS micro data. An estimator for diagonal elements of Σ_{12} is easily shown to be (Angrist, 1989c)

$$\sigma_{12j} = \left[p_j^l\left(L_j - \mu_j^c\right)\right]/n_j.$$

A simplified procedure, described in Section 5 of this Appendix, is used to estimate the nonzero off-diagonal elements of Ω directly from the adjusted cell means.

7. Matching Draft-Eligibility Status to the SIPP

The Survey of Income and Program Participation is a Census Bureau longitudinal survey of approximately 20,000 households in the civilian noninstitutional population (U.S. Department of Commerce, 1985). Data for the first wave of the first SIPP panel were collected from four rotation groups in 1983 and 1984.

The SIPP public-use tapes contain year and month of birth. A Census Bureau "in-house" version of the

[20] This formula ignores the fact that the censored earnings of men with multiple employers may be above the taxable maximum. The formula may be used to analyze the bias in treatment effects estimated from censored data. Suppose that cell j is for a sample of draft-eligible men and that cell k is for a sample of draft-ineligible men. The draft-eligibility treatment effect estimated from the difference between CWHS censored mean earnings in cells j and k is the sample analogue of

$$\mu_j^c - \mu_k^c = \left(\mu_j^0 - \mu_k^0\right) + \left(p_j^l - p_k^l\right)\left(L_j - \mu_j^l\right) + p_k^l\left(\mu_k^l - \mu_j^l\right).$$

Assuming that the effect of draft eligibility on both p^l and μ^l is of the same sign as the effect on μ^0, this expression shows that treatment effects estimated from the censored data differ from the true treatment effect by terms that are opposite in sign from the true effect.

SIPP contains information on day of birth, which is not released to the public. At the author's request, this information was used to match a dummy variable for draft-eligibility status to the public use version of SIPP Panel I. Draft eligibility was determined by the official RSN ceiling for men born from 1944–52 and by RSN 95 for men born in 1953. Vietnam era veteran status is coded from the SIPP variables VETSTAT, which records veteran status, and from U-SRVDTE, which records the period of service.

REFERENCES

Angrist, Joshua D., "Grouped Data Estimation and Testing in Simple Labor Supply Models," Industrial Relations Section Working Paper, no. 234, Princeton University, July 1988.

______, (1989a), "Using the Draft Lottery to Measure the Effects of Military Service on Civilian Labor Market Outcomes," in *Research in Labor Economics*, *10*, Ron Ehrenberg, ed., Greenwich: JAI Press, 1989.

______, (1989b), "Selection for Military Service in the Vietnam Era," Industrial Relations Section Working Paper, no. 250, Princeton University, April 1989.

______, (1989c), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," Industrial Relations Section Working Paper, no. 251, Princeton University, April 1989.

______, (1989d), "Instrumental Variables Regression with Moments from Two-Samples," Chapter 3 in *Econometric Analysis of the Vietnam Era Draft Lottery*, unpublished Princeton University doctoral dissertation, October 1989.

______ **and Krueger, Alan**, "Why Do World War II Veterans Earn More?" National Bureau of Economic Research Working Paper, no. 2991, May 1989.

Baskir, Lawrence M. and Strauss, William A., *Chance and Circumstance: The Draft, The War and the Vietnam Generation*, New York: Alfred A. Knopf, 1978.

Berger, Mark C. and Hirsch, Barry T., "The Civilian Earnings Experience of Vietnam-Era Veterans," *Journal of Human Resources*, Fall 1983, *18*, 455–79.

Buckler, Warren and Smith, Cresston, "The Continuous Work History Sample (CWHS): Description and Contents," in *Statistical Use of Administrative Records: Recent Research and Present Prospects Volume I*, Department of the Treasury, Internal Revenue Service, Statistics of Income Division, March 1984.

Card, David and Sullivan, Daniel, "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment," *Econometrica*, January 1988, *56*, 497–530.

Chamberlain, Gary, "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, January 1982, *18*, 5–46.

Crane, Jon R. and Wise, David A., "Military Service and the Civilian Earnings of Youths," Chapter 6 in *Public Sector Payrolls*, D. A. Wise, ed., Chicago: University of Chicago Press, 1987.

DeTray, Dennis N., "Veteran Status as a Screening Device," *American Economic Review*, March 1982, *72*, 133–42.

Griliches, Z. and Mason, William, "Education, Income and Ability," *Journal of Political Economy*, May/June 1972, *80*, Part 2: S74–S103.

Hearst, Norman, Newman, Tom B. and Hulley, Stephen B., "Delayed Effects of the Military Draft on Mortality: A Randomized Natural Experiment," *New England Journal of Medicine*, March 6, 1986, *314*, 620–24.

Heckman, James J. and Robb, Richard, "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, New York: Cambridge University Press, 1985.

Meyer, Robert J., *Social Security*, 3rd ed., Homewood, IL: Richard D. Irwin, 1985.

Millea, Mary T. and Kilss, Beth, "Exploration of Differences Between Linked Social Security and Internal Revenue Service Wage Data for 1972," in *Studies from Interagency Data Linkages*, Report No. 11: Measuring the Impact on Family and Personal Income Statistics of Reporting Differences Between the Current Population and Administrative Sources, U.S. Department of Health, Education and Welfare, Social Security Administration, March 1980.

Mincer, Jacob, *Schooling, Experience and Earnings*, New York: National Bureau of Economic Research, 1974.

Newey, Whitney K., "Generalized Method of Moments Specification Testing," *Journal of Econometrics*, September 1985, *29*, 229–56.

Rosen, Sherwin and Taubman, Paul, "Changes in Life Cycle Earnings: What Do Social Security Data Show?" *Journal of Human Resources*, Summer 1982, *17*, 321–38.

Schwartz, Saul, "The Relative Earnings of Vietnam and Korean-Era Veterans," *Industrial and Labor Relations Review*, July 1986, *39*, 564–72.

Tarr, Curtis W., *By the Numbers: The Reform of the Selective Service System 1970–72*, Washington: National Defense University Press, 1981.

Taussig, Michael K., "The Rationale for Veterans Benefits," in *Those Who Served: Report of the 20th Century Fund Task Force on Policies Toward Veterans*, New York: The 20th Century Fund, 1974.

Wald, Abraham, "The Fitting of Straight Lines If Both Variables Are Subject to Error," *Annals of Mathematical Statistics*, September 1940, *11*, 284–300.

Welch, Finis, "The Effect of Cohort Size on Earnings: The Baby Boom Babies Financial Bust," *Journal of Political Economy*, October 1979, Part 2, *87*, S69–S97.

Willis, Robert J., "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions," in O. Ashenfelter and R. Layard, eds., *The Handbook of Labor Economics*, Amsterdam: North-Holland, 1986.

Council of Economic Advisors, *Economic Report of the President*, February 1988, Washington: USGPO, 1988.

Selective Service System, Office of Public Affairs, "*A Short History of the Selective Service System*," Washington: USGPO, 1986.

______, *Semiannual Report of the Director of Selective Service for the Period July 1 to December 31, 1970*, Washington: USGPO, 1971.

______ (1969–1973), *Semiannual Report of the Director of Selective Service*, selected issues, Washington: USGPO, 1969–73.

U.S. Department of Health and Human Services, Social Security Administration, *Social Security Bulletin Annual Statistical Supplement*, Washington: USGPO, 1987.

U.S. Department of Commerce, Bureau of the Census, *Survey of Income and Program Participation: Wave I Rectangular File*, Washington: USGPO, 1985.

Veterans Administration (1981a), Reports and Statistics Service, Research Monograph 15: *Educational and Income Characteristics of Veterans, March 1979*, Washington: USGPO, August 1981.

______, (1981b), Office of Reports and Statistics, *Data on Vietnam Era Veterans*, Washington: USGPO, September 1981.

______, (1984), Administrator of Veterans Affairs, *Annual Report 1983*, Washington: USGPO, 1984.

Unintended Impacts of Public Investments on Private Decisions: The Depletion of Forested Wetlands

By ROBERT N. STAVINS AND ADAM B. JAFFE*

By affecting relative economic returns, public infrastructure investments can induce major changes in private land use. We find that 30 percent of forested wetland depletion in the Mississippi Valley has resulted from private decisions induced by federal flood-control projects, despite explicit federal policy to preserve wetlands. Our model aggregates individual land-use decisions using a parametric distribution of unobserved land quality; dynamic simulations are used to quantify the impacts on wetlands of federal projects and other factors. (JEL 717)

Private land-use decisions can be affected dramatically by public investments in highways, waterways, flood control, or other infrastructure. The large movement of jobs from central cities to suburbs in the postwar United States and the current destruction of Amazon rain forests have occurred with major public investment in supporting infrastructure. As these examples suggest, private land-use decisions can generate major environmental and social externalities. Hence, the extent to which major investment programs create "secondary impacts" through their effects on private decisions is a matter of great public concern. In this paper, we demonstrate that the depletion of forested wetlands in the Mississippi Valley (an important environmental problem and a North American precursor to the loss of South American rain forests) has been and is currently exacerbated by federal water-project investments, despite explicit federal policy to protect wetlands.

We begin with a structural model of an individual landowner's decision of whether or not to convert a parcel of land from its natural, forested state to agricultural use. This problem can be characterized as an optimal stopping problem, and is similar to those investigated by Glenn Gotz and John McCall (1980 and 1984), Ariel Pakes (1986), John Rust (1987), and Mark Rosenzweig and Kenneth Wolpin (1988). Unfortunately, such models require individual data for estimation; but in the present context data on land-use status and other characteristics of individual parcels would be prohibitively expensive to obtain over an area large enough to contain significant variation in the extent of federal investment. Therefore, we need a model that will explain the proportion of some aggregate (counties, in our case) that will be converted; this leads to the problem of the appropriate aggregation of individual optimal decisions to the county level.

If the land in a county could be assumed to be homogeneous, then the required aggregation could be accomplished simply by modeling a "representative parcel" and assuming that variations within counties were purely random. Such an assumption, however, is untenable. Some land in a county is

*John F. Kennedy School of Government, Harvard University, Cambridge, MA: 02138, and Department of Economics, Harvard University, Cambridge, MA: 02138, respectively. Helpful comments on earlier versions of the paper were provided by: Zvi Griliches, Partha Dasgupta, Christopher Cavanagh, Joseph Kalt, Peter Timmer, Robert Pindyck, Hendrik Houthakker, Mark Watson, Jon Goldstein, Lawrence Goulder, Marie Lynn Miranda, James Hines; seminar participants at Harvard, Stanford, Michigan, Northwestern, the Rand Corporation, Resources for the Future, and the World Resources Institute; and two anonymous referees. The authors alone, however, are responsible for any remaining errors. Financial support was provided by grants from Resources for the Future, the U.S. Department of the Interior, and the Environmental Defense Fund.

of higher (potential) agricultural quality and will therefore be developed first. If we ignore this reality and estimate a model on a panel of counties over time, any variable whose value is initially high, and then falls, will appear to cause conversion. More generally, if we are to predict, based on the past, what would happen if we change policy in the future, it is necessary to take into account the fact that the marginal acre today is different from the marginal acre of the past.

We solve the problem by positing a parametric distribution for the unobserved quality of land within a county. The parameters of this distribution are then estimated jointly with the parameters of the individual-level, structural model. This allows us to quantify the effects of federal programs and policies, via dynamic simulations with the estimated distributional and structural parameters.[1]

The first section of the paper describes the problem of wetland depletion. The second section develops the theoretical framework, including the modeling of unobserved heterogeneity of land parcels. Section III presents econometric results, used in Section IV to carry out dynamic simulations. Concluding comments are found in Section V.

I. Forested Wetlands and Public Policy

Forested wetlands are among the world's most productive ecosystems, providing improved water quality, erosion control, floodwater storage, timber, fish and wildlife habitat, and recreational opportunities. Their continuing depletion is a serious land-use problem; preservation and protection of wetlands have been major federal environmental policy goals for at least twenty years. The largest remaining wetland habitat in the continental United States is the bottomland hardwood forest of the Lower Mississippi Alluvial Plain. Originally covering 26 million acres in seven states, this resource was reduced to about 12 million acres by 1937. Since then, another 6.5 million acres have been cleared, primarily for conversion to cropland.

The owner of a wetland parcel faces an economic decision involving revenues from the parcel in its natural state (primarily from timber), costs of conversion (the cost of clearing the land minus the resulting forestry windfall), and expected revenues from agriculture. Agricultural revenues depend on prices, yields, and, significantly, the drainage and flooding frequency of the land. Needless to say, landowners typically do not consider the positive environmental externalities generated by wetlands; thus conversion may occur more often than is socially optimal.

These externalities are the motivation for federal policy aimed at protecting wetlands, as embodied in the Clean Water Act. Nonetheless, the federal government has engaged in major public investment activities, in the form of U.S. Army Corps of Engineers (Corps) and U.S. Soil Conservation Service (SCS) flood-control and drainage projects, which appear to make agriculture more attractive and thereby encourage wetland depletion. The significance of this effect is disputed by the agencies which construct and maintain these projects; they attribute the extensive conversion exclusively to rising agricultural prices (U.S. Army Corps of Engineers, 1981). Our approach allows us to sort out the effects of federal projects and other changing economic forces. As we will see, these public investments appear to have been a substantial factor causing conversion of wetlands to agriculture.[2]

[1] The problem of estimating a transition process for individuals on the basis of aggregate data arises in many areas. In natural resource economics, in particular, it has been identified as a major gap between the theoretical literature on optimal extraction and the empirical literature on resource depletion (Douglas Bohi and Michael Toman, 1984). Also related is work in demography (for example, Richard Gill, 1986) and technological diffusion (Zvi Griliches, 1957; Paul David, 1966; and Richard Pomfret, 1976).

[2] Theoretical models of wetland conversion are found in John Brown (1972) and Leonard Shabman (1980). Our empirical results should be compared with Randall Kramer and Leonard Shabman 1986, and U.S. Department of the Interior 1988.

II. The Privately Optimal Land-Use Decision and Its Implications for the Behavior of Heterogeneous Aggregates

The first step is the construction of a dynamic optimization model of forestry and agricultural production at the individual, landowner level. The solution of this model yields necessary conditions for conversion of individual forest parcels to agricultural production and for abandonment of parcels of cropland. An explicit model of the heterogeneity of land allows for the aggregation of the respective necessary conditions to the county level, so that an econometrically estimable model can be specified.

A. *A Dynamic Optimization Model of Forestry and Agricultural Production*

Landowners observe a variety of economic, hydrologic, and climatic factors relevant to decisions regarding the use of their lands for forestry or agricultural production. Current and past values of variables presumably constitute the basis for expectations about future values of variables. In particular, landowners observe agricultural prices and production costs, typical agricultural yields for their area, typical timber returns, and the suitability of individual land parcels for agriculture. A prime factor determining such suitability of land (in the geographic area of this study) is its wetness, that is, the degree of (natural and artificial) protection from flooding and poor drainage.

A landowner continually faces a decision of whether to keep land in its current state, to convert forested land to agricultural production,[3] or to abandon cropland and allow it to return to forest. A risk-neutral[4] landowner faced with the decision of how to utilize his land, given the alternatives of forestry and agriculture, may be expected to seek to maximize the present discounted value of the stream of expected future returns to his land:[5]

$$(1)\quad \max_{\{g_{ijt}v_{ijt}\}} \int_0^{\infty} \Big[[A_{it}q_{ijt} - AC_{it}][g_{ijt} - v_{ijt}] - C_{it}g_{ijt} + f_{it}S_{ijt} + w_{it}g_{ijt} - D_{it}v_{ijt} \Big] e^{-r_t t}\, dt$$

subject to:

$$(2)\quad \dot{S}_{ijt} = v_{ijt} - g_{ijt}$$

$$(3)\quad 0 \le g_{ijt} \le \bar{g}_{ijt}$$

$$(4)\quad 0 \le v_{ijt} \le \bar{v}_{ijt}$$

where i indexes counties, j indexes individual land parcels, and t indexes time; uppercase letters represent stocks or present values; and lowercase letters represent flows. The variables are

A = discounted present value of the infinite stream[6] of typical expected agricultural revenues per acre in the county;

[3] The potential sale of a parcel is economically irrelevant, since a new owner faces the same conversion/abandonment decision. The option of residential or commercial development is empirically insignificant. Finally, since land prices presumably reflect the present value of net revenues under optimal use, land-price data, if available, would provide an alternative means of examining the phenomena modeled here. See Richard Arnott and Frank Lewis (1979).

[4] Evidence on the risk aversion of farmers is not consistent (Douglas Young 1979; Bruce Gardner and Jean-Paul Chavas 1979; Rulon Pope 1981). Provision for risk-averse behavior in the objective functional would lead to the inclusion in derived necessary conditions of second-order moments of stochastic variables. Due to lack of sufficient data, however, only expected values are used in the empirical analysis: we assume risk neutrality and independence of relevant factors. Because flood protection projects may reduce the variance of returns (in addition to increasing average returns), the assumption of risk neutrality may lead to underestimation of the impacts of federal projects.

[5] Note that the term in the objective function which represents the (discounted present value of) expected future net revenue from agricultural production, $A_{it}q_{ijt} - AC_{it}$, is the price of farmland in a competitive market.

[6] Though a discrete-time formulation would be more realistic, the continuous-time approach is simpler and easier to interpret. The econometric specification implied by the discrete-time formulation is identical.

q = parcel-specific index of feasibility of agricultural production, including effects of soil quality and soil moisture;
g = acres of land converted from forested to agricultural use;
v = acres of cropland abandoned (gradually returned to forested condition);
AC = expected costs of agricultural production per acre, expressed as the discounted present value of an infinite future stream;
C = average cost of conversion per acre (indexed by weather conditions);[7]
f = expected annual net income from forestry per acre;
S = stock (acres) of forests;
r = real interest rate;
W = windfall of net revenue per acre from a one-time clearcut of forest (in the process of conversion);
D = expected present discounted value of loss of income due to the gradual regrowth of forest (harvesting does not occur until the year $t+R$, where R is the exogenously determined rotation length);[8]
$\bar{g}$ = maximum feasible rate of conversion, defined such that

$$\int_{t}^{t+\Delta}\left[\bar{g}_{ij\tau}\right]d\tau=S_{ijt} \tag{5}$$

for arbitrarily small interval, Δ, over which $\bar{g}_{ij\tau}$ is constant;
$\bar{v}$ = maximum feasible rate of abandonment, defined such that

$$\int_{t}^{t+\Delta}\left[\bar{v}_{ij\tau}\right]d\tau=T_{ijt}-S_{ijt} \tag{6}$$

$$=AG_{ijt}$$

for arbitrarily small interval, Δ, over which $\bar{v}_{ij\tau}$ is constant.
AG = stock (acres) of agricultural land; and
T = total acreage of parcel in the flood plain available for conversion.[9]

Note that only the control variables, the state variables, and the quality index, q, are specific to the individual land parcel. All of the revenue and cost variables are measured at the county level. This is a consequence of the data, and is indicative of the information available to landowners. Aggregation of first-order conditions for individual landowners to the county level will yield relationships among county-level variables and the distribution of q. The parameters of these relationships and of the underlying distribution can then be estimated econometrically.

The solution to the optimization problem is provided in a longer version of this paper (Stavins and Jaffe, 1988). To characterize that solution, the following notation is convenient:

$$X_{ijt}=A_{it}\cdot q_{ijt}-AC_{it}-C_{it}-FN_{it}, \tag{7}$$

$$Y_{ijt}=\tilde{F}_{it}-A_{it}\cdot q_{ijt}+AC_{it}, \tag{8}$$

where FN_{jt} is net forestry revenue, $F_{it}-W_{it}$; and $\tilde{F}_{it}$ is delayed net forest revenue, $F_{it}-D_{it}$.

[7] Precipitation and consequent soil moisture are later allowed to influence conversion costs; the conversion cost term in equation (1) is then replaced by $C_{it}\cdot\exp\{\alpha_2 PHDI_{it}\}\cdot g_{it}$, where α_2 is an estimated parameter and $PHDI_{it}$ is the Palmer Hydrological Drought Index.

[8] The inclusion of this term allows for a category of land which is neither productive farm nor mature forest, but evolving "bush." The expected present discounted value of loss of income due to gradual regrowth of forest, D_{it}, is:

$$D_{it}=\int_{t}^{t+R}\left\{f_{i\tau}e^{-r(\tau-t)}\right\}d\tau=F_{it}\cdot\left\{1-e^{-rR}\right\},$$

where F_{it} is the present discounted value of an infinite future stream of net forest income, i.e. $F_{it}\cdot r=f_{it}$. If $R=0$, $D_{it}=0$ (if regrowth is instantaneous, there is no loss of revenue due to harvest delay); and if $R=\infty$, $D_{it}=F_{it}$ (if the regrowth period is infinitely long, there is a complete loss of all forest revenue.)

[9] Some land in the thirty-six counties was withdrawn from availability to the private market during the study period as a result of designation of protected status by Federal and state authorities, including the U.S. Fish and Wildlife Service, the U.S. Forest Service, and state fish and game agencies.

The solution to the maximization problem implies that conversion or abandonment will occur under the following conditions:

(9) Conversion occurs if $X_{ijt} > 0$

and parcel is forested,

(10) Abandonment occurs if $Y_{ijt} > 0$

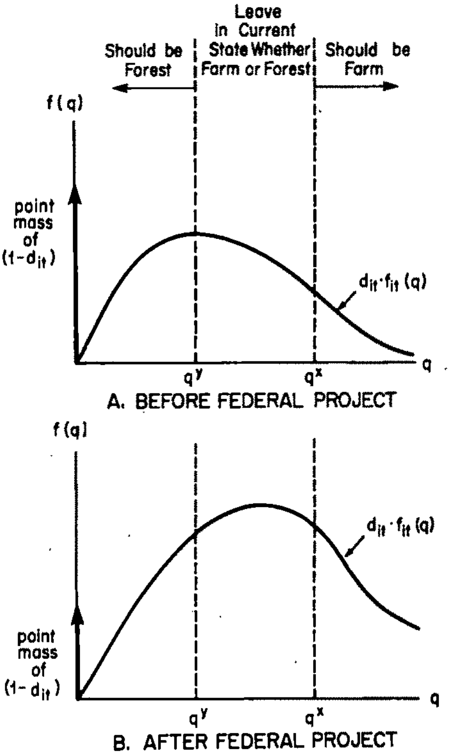
and parcel is cropland.

Letting the time interval Δ in equations (5) and (6) be equal to unity (a single time period), the continuous-time model yields the following result for a discrete-time situation: when conversion occurs, $g^*_{ijt} = S_{ijt}$; and when abandonment occurs, $v^*_{ijt} = AG_{ijt}$. For each homogeneous parcel j, it is always optimal either to convert the entire parcel from forested condition to agricultural use (only, of course, if it is in a forested state), to abandon the entire parcel (only if it is agricultural cropland), or to do nothing.

B. *Modeling the Unobserved Heterogeneity of Land*

Equations (9) and (10) imply that all land in a county of a given quality will be in the same use in the steady state. We do not, however, observe counties as all forest or all farmland. Partly, this may reflect deviations from the steady state (on which more below), but to a great extent it reflects heterogeneity of land within a county. We can characterize this heterogeneity in terms of a probability density function $f_i\{q_{ijt}\}$, as pictured in Figure 1.

The function pictured has a point mass of probability at $q = 0$, corresponding to land of hydrologic condition which renders agriculture completely infeasible. The rest of the land in a county could be farmed, with agricultural yields being an increasing function of q, since q reflects primarily variations in frequency of flooding, drainage, and soil conditions. Thus it depends on the natural lay of the land, the type of soil, and importantly, the existence of man-made flood-control and drainage projects.



A. BEFORE FEDERAL PROJECT

B. AFTER FEDERAL PROJECT

FIGURE 1. THE EFFECT OF FEDERAL FLOOD CONTROL AND DRAINAGE PROJECTS ON THE DISTRIBUTION OF UNOBSERVED LAND QUALITY

The two panels of Figure 1 demonstrate the effects of public projects. Conceptually, there are three such effects, but two are indistinguishable empirically. First, flood-control and drainage projects may render feasible for agriculture land that was previously infeasible; this is reflected as a transfer of probability mass from the point at zero to the rest of the distribution. Second, by improving drainage or reducing flooding, projects may shift the quality distribution for previously feasible land to the right.

Finally, projects may also change the distribution of q for agriculturally feasible land because the land which *was* infeasible that is rendered feasible may have an underlying or potential quality distribution that is different from the distribution for already feasible land. This effect could go in either direction. On the one hand, it could be that infeasible land, because it is low-lying and poorly drained, is likely to remain low quality even if given enough flood protection to be technically feasible. Alternatively, it could be that the soil on such land is of high quality and that, with flood protection, it is very good agricultural land. In the former hypothetical, projects would shift the quality distribution for feasible land to the left; in the latter to the right. Note that if *any* of the infeasible land is potentially of high quality, and if this is known in advance, then the government could increase the tendency for projects to be quality-improving by choosing preferentially to give flood protection to (potentially) high quality land. Thus, both the second and third effects are likely to result in an improvement in observed quality of feasible land, but these effects cannot be distinguished in the data.

We parameterize this model of the quality distribution as follows:[10]

$$\text{(11)}\quad \log(q_{ijt}) \sim N(\mu, \sigma^2) \quad \text{with probability } d_{it}$$

$$q_{ijt} = 0 \text{ with probability } (1 - d_{it})$$

where d_{it} is the probability that agricultural production is feasible.[11] The first effect identified above is incorporated by allowing d_{it} to be a function of the extent of federal projects:

$$\text{(12)}\quad d_{it} = \left[\frac{1}{1 + \left[\frac{1}{e^{\pi(z)}}\right]}\right]$$

$$\text{(13)}\quad \pi(z) = DRY_i + \beta_1 \cdot PROJ_{it}$$

where DRY_i is a measure of the percentage of county i which is naturally protected from periodic flooding, $PROJ_{it}$ is an index of the share of county i at time t which has been artificially protected from flooding by Corps and SCS projects, and β_1 is a parameter which indicates the impact of artificial flood protection relative to the impact of natural flood protection.[12]

The two effects of projects on the quality distribution for feasible land are incorporated by allowing the parameters of the lognormal distribution to depend on the project index as well:

$$\text{(14)}\quad \log(q_{ijt}) \sim N\left[\mu(1 + \beta_2 PROJ_{it}),\ [\sigma \cdot (1 + \beta_3 PROJ_{it})]^2\right] \text{ with prob. } d_{it}$$

$$q_{ijt} = 0 \text{ with probability } (1 - d_{it})$$

The effects of federal projects on land-use decisions are thus captured through the employment of three project-impact parameters —β_1, β_2, and β_3.

C. *Aggregation of Necessary Conditions for Forested Wetland Conversion*

Having posited the basic nature of the heterogeneity of land, the distributional model can now be used to aggregate the individual-landowner necessary conditions previously developed.[13] Equation (9), above, indicates that there is an incentive to convert forested wetlands to agricultural cropland if $X_{ijt} > 0$. Hence, there is a threshold value of q_{ijt}, denoted q_{it}^x, above which the incentive

[10] We focus on the lognormal distribution because we believe that its general shape is appropriate for a distribution over quality of land. We experimented with other functional forms, and we discuss some of these results below.

[11] In the current application, individual county means and variances are not estimated. Note, however, that separate μ_i and σ_i parameters are identified and could therefore, in principle, be estimated, given sufficient data.

[12] The logistic specification is used to constrain d_{it} to values between zero and unity, because the empirical measures of Corps and SCS project impact areas and natural flood protection are only indexes of protection.

[13] One alternative research strategy would be to collect data on individual land parcels and estimate a model such as that developed by Rust (1987). In the present context, however, it would be prohibitively expensive to develop this data over an area large enough to identify the effects of interest.

for conversion manifests itself:

$$(15)\qquad q^x_{it} = \left[\frac{C_{it} + FN_{it} + AC_{it}}{A_{it}}\right].$$

If conversion cost is allowed to be heterogeneous across land parcels (within counties) and flood-control projects are believed to affect conversion costs as well as agricultural feasibility (yields), then the conversion cost term in equation (1) is replaced by the term $\alpha_1 \cdot q_{ijt} C_{it} g_{it}$, where α_1 is a parameter which captures the relative effect of heterogeneity on conversion costs, compared with the effect on agricultural yields. Next, allowing for the parametric effect of weather on conversion costs, q^x_{it} becomes

$$(16)\qquad q^x_{it} = \left[\frac{FN_{it} + AC_{it}}{A_{it} - \alpha_1 C_{it} \cdot \exp\{\alpha_2 PHDI_{it}\}}\right].$$

In either case, there is an incentive to convert parcel j (in county i at time t) from a forested condition to agricultural cropland if $q_{ijt} > q^x_{it}$. Therefore, the privately optimal (the desired or target) stock of converted land, expressed as a fraction of all land available for conversion, is

$$(17)\qquad \left[\frac{AG}{T}\right]^*_{it} = \left[1 - \left[\frac{S}{T}\right]^*_{it}\right] = d_{it} \cdot \left[\int_{q^x_{it}}^{\infty} [f_i\{s\}]\, ds\right],$$

where $f_i\{\cdot\}$ is the lognormal density function. Therefore,

$$(18)\qquad \left[\frac{AG}{T}\right]^*_{it} = d_{it} \cdot [1 - F_i[q^x_{it}]],$$

where $F_i[\cdot]$ is the lognormal cumulative distribution function, and

$$(19)\qquad \left[\frac{AG}{T}\right]^*_{it} = d_{it} \cdot [1 - \mathbf{F}[[\log[q^x_{it}] - \mu]/\sigma]],$$

where $\mathbf{F}[\cdot]$ is the cumulative, standard normal distribution function.

There is an analogous equation for abandonment, which gives the target stock of forested land as a fraction of the total

$$(20)\qquad \left[\frac{S}{T}\right]^*_{it} = d_{it} \cdot [\mathbf{F}[[\log[q^y_{it}] - \mu]/\sigma]] + [1 - d_{it}],$$

where q^y_{it} is the threshold value of q_{ijt} below which the incentive for abandonment manifests itself (see equation (8), above, and Table 1).

These relationships are shown graphically in Figure 1. The area to the right of q^x should be converted to farmland (if forested); the area to the left of q^y should be abandoned (if farmed). The area between q^y and q^x is worth farming if previously converted, but not worth converting if currently forested. Changes in the economic climate affect the position of the thresholds. For example, an increase in expected agricultural revenue will shift both the conversion threshold, q^x, and the abandonment threshold, q^y, to the left, thereby encouraging conversion of forested land to agricultural use.

III. Econometric Analysis

A. *Specification Issues*

Two specification issues must be addressed before the model embodied in equations (19) and (20) can be estimated. These are: (1) the possibility that adjustment toward optimal land use is not instantaneous; and (2) combining the conversion and abandonment models into one estimating equation.

All of the analysis to this point has assumed that conversion to optimal land use (conditional on current prices) occurs instantaneously. There are several reasons why this may not be true.[14] Although we estimate the

[14] These include: forest age distribution, liquidity constraints, uncertainty about the permanence of price movements, and decision-making inertia.

D 5770

TABLE 1—ECONOMETRIC MODEL OF FORESTED WETLAND CONVERSION AND AGRICULTURAL CROPLAND ABANDONMENT

$$FORCH_{it} = FORCH^c_{it} \cdot D^c_{it} + FORCH^a_{it} \cdot D^a_{it} + \lambda_i + \phi_{it}$$

$$\begin{aligned} FORCH^c_{it} \cdot (-1) = \gamma_c\Big[d_{it} \cdot \Big[1 - \mathbf{F}\Big[\big[\log[q^x_{it}] - \mu(1+\beta_2 PROJ_{it})\big] \\ /\sigma(1+\beta_3 PROJ_{it})\Big]\Big] + \left.\frac{S}{T}\right]_{i,t-1} - 1\Big] \end{aligned}$$

$$\begin{aligned} FORCH^a_{it} = \gamma_a\Big[d_{it} \cdot \Big[\mathbf{F}\Big[\big[\log[q^y_{it}] - \mu(1+\beta_2 PROJ_{it})\big] \\ /\sigma(1+\beta_3 PROJ_{it})\Big]\Big] + [1 - d_{it}] - \left[\frac{S}{T}\right]_{i,t-1}\Big] \end{aligned}$$

$$d_{it} = \left[\frac{1}{1 + \left[\dfrac{1}{e^{\pi(z)}}\right]}\right], \text{ where } \pi(z) = DRY_i + \beta_1 PROJ_{it}$$

$$q^x_{it} = \left[\frac{FN_{it} + AC_{it}}{A_{it} - \alpha_1 C_{it} \cdot \exp\{\alpha_2 PHDI_{it}\}}\right] \qquad q^y_{it} = \left[\frac{\tilde{F}_{it} + AC_{it}}{A_{it}}\right].$$

"frictionless" model implied by equations (19) and (20), we also want to consider the possibility of partial adjustment in each observation period toward the optimal land use pattern. In the case of conversion, we have

$$(21) \quad \left[\frac{AG}{T}\right]_{it} + \left[\frac{AG}{T}\right]_{i,t-1} = \gamma_c \cdot \left[\left[\frac{AG}{T}\right]^*_{it} - \left[\frac{AG}{T}\right]_{i,t-1}\right] + \varepsilon^c_{jt}$$

where γ_c is the rate of partial adjustment, $[AG/T]^*$ is given by equation (19), and ε^c_{it} is an error term composed of a county-specific (time-invariant) component, λ_i, and a component, ϕ^c_{it}, which has mean zero, so that $\varepsilon^c_{it} = \lambda_i + \phi^c_{it}$. In abandonment situations, we have

$$(22) \quad \left[\frac{S}{T}\right]_{it} - \left[\frac{S}{T}\right]_{i,t-1} = \gamma_a \cdot \left[\left[\frac{S}{T}\right]^*_{it} - \left[\frac{S}{T}\right]_{i,t-1}\right] + \varepsilon^a_{it},$$

where γ_a is the rate of partial adjustment for abandonment, $[S/T]^*$ is given by equation (20), and ε^a_{it} is an error term composed of a county-specific component, λ_i, and a component, ϕ^a_{it}, which has mean zero. Since county-level stocks are aggregates of individual decisions, these adjustment parameters represent the probability that a landowner not in equilibrium in a given time period will switch to the optimal land use within the initial period.[15]

To combine equations (21) and (22) into one relationship, we define the net change in the forested fraction of the county between

[15] It might seem that a superior approach would be to incorporate adjustment costs or lags into the original optimization problem, but this cannot be done in a way which yields necessary conditions which can be aggregated across heterogeneous parcels to the county level. Any such mechanism must depend on deviations of *individual* parcels from optimality. Estimating a model with adjustment costs requires observing the relationship between the magnitude of deviations from equilibrium and the rate of movement. Since we do not observe individual parcels, this cannot be done, so any adjustment mechanism built into the *individual* model could not be estimated with county data. One could specify a version of equation (1) with adjustment costs at the county level, but that would be equivalent to a representative-firm assumption. Thus, a fully dynamic optimal model can only be implemented with individual data.

periods t-1 and t as

$$(23)\quad \left[\frac{AG}{T}\right]_{it} - \left[\frac{AG}{T}\right]_{i,t-1} = \left[\frac{S}{T}\right]_{i,t-1} - \left[\frac{S}{T}\right]_{it} = (-1)\cdot FORCH_{it}.$$

Under the assumptions of the model, conversion and abandonment will never occur simultaneously in the same county, so we can write

$$(24)\quad FORCH_{it} = -D_{it}^{c}\gamma_{c}\left[\left[\frac{AG}{T}\right]_{it}^{*} - \left[\frac{AG}{T}\right]_{i,t-1}\right] + D_{it}^{a}\gamma_{a}\left[\left[\frac{S}{T}\right]_{it}^{*} - \left[\frac{S}{T}\right]_{i,t-1}\right] + \varepsilon_{it},$$

where D_{it}^{c} and D_{it}^{a} are dummy variables[16] for the conversion and abandonment regimes; $[AG/T]^{*}$ and $[S/T]^{*}$ are the corresponding target stocks; and ε_{it} is a composite error term, defined by $\varepsilon_{it} = \varepsilon_{it}^{c} + \varepsilon_{it}^{a} = \lambda_{i} + \phi_{it}^{c} + \phi_{it}^{a} = \lambda_{i} + \phi_{it}$.

The county-specific components of the error term, λ_{i}, are treated as fixed-effect parameters and the ϕ_{it} are assumed to be independently distributed across i and t, but not necessarily homoscedastic.[17] Thus, equation (24) is a single-equation, fixed-effects model, the parameters of which can be estimated by nonlinear least squares, with county dummy variables employed to eliminate any bias due to the county fixed effect (Table 1).

B. *Parameter Estimation of Alternative Specifications*

Using data for 36 counties in Arkansas, Louisiana, and Mississippi, during the period 1935–1984,[18] the parameters of the model embodied in equations (24), (Table 1) were estimated econometrically. Panel data were incorporated into the estimation process by stacking the data for 36 counties for each of ten (five-year[19]) time periods, for a total of 360 observations.

The results of six versions of the model are presented in Table 2. The overall results lend support to the basic validity of the model. Estimated parameters are all of the expected sign, and nearly all estimates are significant at the 90 percent, 95 percent, or 99 percent level. Also, both parameter and standard error estimates are quite robust with respect to modifications of the specification. Thus, the basic structural model of changes in forested acreage being a function of expectations regarding relative economic returns from agriculture and forestry is strongly supported.[20] In addition, the fixed-effects approach is clearly superior to a totally pooled model, as indicated by the appropriate likelihood ratio tests.[21]

Column 1 of Table 2 restricts flood-control and drainage projects to affecting only agricultural feasibility, while columns 2 and 3 also allow for effects on quality. The estimated partial adjustment coefficients on conversion, γ_{c}, and abandonment, γ_{a}, indicate that about 60 percent of the targeted de-

[16] The dummies are endogenous variables. We first estimate separately the conversion and abandonment equations, and thus predict values for conversion and abandonment. The two equations are then combined as in equation (24) with dummies constructed on the basis of the predicted conversion or abandonment.

[17] Heteroscedasticity-consistent standard errors were calculated according to Halbert White (1980), and are reported in the table of econometric results. The possibility of serial correlation in ϕ_{it} was explored. Neither J. Durbin's (1970) test nor that suggested by T. S. Breusch (1978) and L. G. Godfrey (1978) indicated significant serial correlation.

[18] The nature and sources of data employed are briefly described in the Appendix, which also provides basic statistics for all variables.

[19] Limitations on the availability of data on the dependent variable (forested acreage) necessitated the use of a quinquennial as opposed to an annual model.

[20] The possibility exists of including in the model a measure of individuals' expectations regarding future construction of flood-control and drainage projects, proxied by project authorizations. Given relatively constant real conversion costs, however, no incentive exists for landowners to convert their parcels prior to project construction (and consequent flood protection).

[21] For example, in the $L1$ specification, the likelihood ratio (LR) statistic is 69.9; the appropriate χ^2 critical value is 58.6 at the 99 percent confidence level.

TABLE 2—ECONOMETRIC ESTIMATION RESULTS—LOGNORMAL, NORMAL, AND UNIFORM DISTRIBUTIONS OF LAND QUALITY

	Alternative Specifications[a]					
Parameter	$L1$	$L2$	$L3$	$L4$	$N1$	$U1$
γ_a Abandonment	0.37618	0.32360	0.36717	–	0.41883	0.18288
Partial Adjustment	(0.190)[b]	(0.177)	(0.184)		(0.190)	(0.075)
γ_c Conversion	0.44875	0.69352	0.64826	–	0.62814	0.29872
Partial Adjustment	(0.142)	(0.156)	(0.154)		(0.150)	(0.114)
μ Mean of Unobserved	0.74095	0.83464	1.11650	1.41950	2.26650	–
Quality Distribution	(0.368)	(0.290)	(0.364)	(0.354)	(0.419)	
σ Standard Deviation	0.38182	0.44438	0.43848	0.56324	0.43538	–
of Quality Distrib.	(0.087)	(0.069)	(0.067)	(0.021)	(0.067)	
ω Upper Limit of	–	–	–	–	–	4.64270
Quality Distribution						(2.173)
θ Range of Unobserved	–	–	–	–	–	1.34980
Quality Distribution						(0.855)
β_1 Project Impact on	9.20170	8.83060	8.93700	8.37430	8.69140	8.94940
Agric. Feasibility	(3.216)	(2.309)	(2.465)	(1.768)	(2.394)	(3.705)
β_2 Project Impact on	–	1.07240	0.77193	0.36821	0.24691	–
Heterogeneity Mean		(1.467)	(0.774)	(0.449)	(0.317)	
β_3 Project Impact on	–	0.53757	0.42799	0.36451	0.39361	–
Heterogeneity S.D.		(0.229)	(0.183)	(0.133)	(0.176)	
α_1 Relative Conversion	1.58160	1.02070	–	–	–	–
Cost Impact	(0.923)	(1.169)				
α_2 Weather Impact on	–	–	1.59720	1.59410	1.41720	1.58600
Conversion Cost			(0.304)	(0.296)	(0.193)	(0.302)
Goodness-of-Fit[c]	0.6719	0.6743	0.6747	0.6681	0.6738	0.6742
Log Likelihood Value	786.07	790.62	791.70	788.23	791.57	787.89
Degrees of Freedom	318	316	316	318	316	318

[a] All versions also contain 36 county dummies. $L1$, $L2$, $L3$, and $L4$ (frictionless model) employ lognormal distributions of land quality; $N1$ employs a normal distribution; and $U1$ employs a uniform distribution.

[b] Robust standard error estimates appear below parameter estimates.

[c] This dynamic goodness-of-fit statistic is equal to one minus Theil's U-statistic, based on comparing predicted and actual net rates of conversion.

crease in forested acreage (increase in agricultural cropland) and about 36 percent of the targeted abandonment occur in the initial five-year period.

The distribution of heterogeneity is nondegenerate: the standard deviation, σ, of the lognormal distribution of agricultural quality is significant quantitatively and statistically. Likewise, two of the three parameters capturing the impact of federal projects on conversion and abandonment are significant: direct impact on agricultural feasibility, β_1; and impact on the standard deviation of agricultural quality, β_3. The impact on the mean of agricultural quality, β_2, is positive but not significant.[22] The average direct impact of artificial flood protection on agricultural feasibility, β_1, is about nine times that of "natural flood protection."

[22] Note that the model is parameterized such that projects affect the mean and variance of the log of q. The expectation of q itself is

$$\exp\{\,\mu(1+\beta_2 PROJ)+0.5\sigma^2(1+\beta_3)^2\},$$

so both β_2 and β_3 indicate increasing agricultural quality due to projects.

It was not possible, due to limitations of the data, to estimate the equation with both a parameter for the effect of conversion costs relative to other benefits and costs, α_1, and a parameter for the effect of weather on conversion costs, α_2, although the parameters of such a specification are theoretically identified. The specification with α_1 ($L2$) appears inferior to the one with α_2 ($L3$), and the estimate of α_1 is not significantly different from 1.0. The impact of weather on conversion costs, α_2, is very significant.

Columns $L4$, $N1$, and $U1$ explore the sensitivity of the results to dynamic and distributional assumptions. $L4$ assumes instantaneous rather than partial adjustment to the optimal state. This constraint is rejected by the data,[23] but most parameters remain qualitatively similar. The dynamic goodness-of-fit, calculated as suggested by Henri Theil (1961), shows only slight decline compared to the partial adjustment case ($L3$).[24] Assuming a normal ($N1$) or uniform ($U1$) distribution of unobserved land quality also yields results that are qualitatively similar, with slight decrease of dynamic goodness-of-fit.

Although the results in Table 2 exhibit the significance of prices, costs, and government projects in conversion and abandonment decisions, these results say little about the relative importance of these influences. Due to the nonlinear, dynamic form of the model, the importance of the various factors can be discerned only through a series of dynamic simulations.

IV. Dynamic Simulation Results

To provide a benchmark for comparison, the extent of conversion or abandonment is simulated using the econometrically estimated parameters and the actual, historical values of all variables ("factual simulation"). Then, in a series of "counterfactual simulations," the extent of conversion or abandonment is simulated using various assumed counterfactual values for certain exogenous variables, while maintaining all other variables at their actual levels. Finally, the simulated changes in forested wetland acreage in each counterfactual simulation are compared to the factual simulation changes. Any difference represents an estimate of the land-use change that can be attributed to a change in an exogenous variable from the counterfactual value of the variable to its actual historical pattern of values.

The simulation results utilizing model $L3$ are summarized in Table 3. Column 1 shows the total (net) conversion of forested wetlands to farmland through 1984. In the factual simulation, this is 3.68 million acres. (The true historical conversion was 3.64 million.) Because of the partial adjustment mechanism, conversion will continue into the future, even if all exogenous variables remain unchanged. To capture this effect, column 2 extends the simulations through 1999, keeping all (factual) variables at their 1984 values. Thus, we predict that a total of 3.83 million would be converted by the end of the century if 1984 conditions were to prevail. The target stock based on 1984 values corresponds to net conversion of 3.84 million acres, so the 1999 simulations come very close to the steady state.

The second row in the table shows that simulated wetland depletion through 1999 if *no* federal projects had been built is about 32 percent less than factually simulated depletion. For comparison, the third row shows simulated depletion if flood protection provided by natural topography and the Missis-

[23] Individually, the Wald statistics for constraining the model to instantaneous adjustment are 5.2 for conversion and 11.8 for abandonment, compared with a 99 percent critical (χ^2) value of 6.6; the joint test yields a statistic of 12.1, compared with a 99 percent critical value of 9.2.

[24] A frequently used measure of dynamic performance is the root-mean-squared (RMS) error, but this measure suffers from the limitation that its magnitude is not standardized. An alternative is Theil's inequality coefficient (Theil, 1961). The numerator of this statistic is the RMS error, and the scaling of its denominator ensures that values fall within the bounds of zero and unity, where zero indicates a perfect dynamic fit. In keeping with the ordering of most goodness-of-fit measures, the final comparative statistic shown in Table 2 is equal to one minus the Theil coefficient, so that a perfect fit is evidenced by a value of 1.0.

TABLE 3—SIMULATED CHANGE IN STOCK OF FORESTED WETLANDS UNDER ALTERNATIVE SCENARIOS LOGNORMAL DISTRIBUTION OF LAND QUALITY

	Change[a] in Forested Wetland Acreage		65-Year Period 1935–1999	
Simulation	50-Year Period 1935–84	65-Year Period 1935–99	Percentage of Factually Simulated Depletion	Share of Depletion Due to This Factor[b] in percent
	(1)	(2)	(3)	(4)
(1) Factual	−3.677	−3.834	(100)	(0.0)
(2) No Federal Flood-Control or Drainage Projects	−2.527	−2.612	68.1	31.9
(3) No Flood Protection from Natural Topography and Mainline Levees	−2.831	−2.984	78.8	21.2
(4) Conversion Costs Set to Zero	−4.354	−4.526	118.0	−18.0
(5) Net Forestry Revenue Set to Zero	−4.259	−4.419	115.3	−15.3
(6) Agricultural Prices Kept at 1934 Levels	−3.641	−3.818	99.6	0.4

[a]Millions of acres, based upon parameter estimates from specification $L3$, reported in column 3 of Table 2.
[b]Difference between counterfactual simulation and factual simulation, divided by factual simulation.

sippi mainline levee system[25] were eliminated. This has less effect than elimination of federal projects, reducing depletion by 21 percent. Rows 4 and 5 show that conversion costs and forestry revenues were significant forces restraining conversion. Simulated conversion is 18 percent more if conversion is costless, and 15 percent more if forestry yielded no net revenue.

Finally, the last row of Table 3 explores the hypothesis (maintained by the Corps of Engineers and others) that rising agricultural prices drove wetland depletion, by simulating conversion if farm prices had held at their 1934 levels (in real terms). Net depletion is about 1 percent less than when factually simulated. Thus, there is no evidence that rising agricultural prices were a significant factor driving conversion. Even with the depressed farm prices of 1934, economic incentives favored conversion of many acres.

The simulations in Table 3 were also carried out with the frictionless, normal, and uniform models ($L4$, $N1$, and $U1$ in Table 2). Although the changes attributed to specific factors vary somewhat, the ranking of factors by importance is the same in all versions. Two differences, however, are worth noting. The frictionless model, by assumption, does not predict continued conversion after 1984 based on 1984 conditions. In fact, it predicts slightly less net conversion by 1984 (3.586 million acres). As noted above, the optimal stock corresponds to net conversion of 3.84 million acres, so the partial adjustment model run out to 1999 would appear to be a better indicator of the ultimate conversion. At the other extreme, the uniform model predicts net conversion of 4.065 million acres by 1999, more than twice

[25]"Effect of natural topography" refers to DRY_i ($=1 - FLRISK_i$), a measure of natural flood-proneness of counties (prior to construction of flood-control and drainage projects). The "mainline levee system" (MLS) along the Mississippi River, which was in place virtually from the beginning of the study period, may also have had an impact on wetland depletion. Data on the area protected by the MLS are not available. The protected area is approximately proportional to the area affected by the great flood of 1927, but flood data are highly correlated with natural topography, and so multicollinearity prohibits estimation of the effects of natural topography and the MLS in the same equation. With the omission of the MLS proxy variable, the $FLRISK$ variable accounts for both phenomena. Thus, the reported impact of Federal projects on conversion refers exclusively to interior levee development and underestimates the impact of all projects.

as much conversion between 1984 and 1999 as is simulated using the preferred, lognormal model. Not surprisingly, the impact of distributional assumptions becomes greater as we move into the tails of the distributions.

V. Conclusions and Policy Implications

The statistical analysis leads to several conclusions. First, landowners responded to economic incentives in their land-use decisions. Second, construction of federal flood-control and drainage projects caused a higher rate of conversion of forested wetlands to croplands than would have occurred in the absence of projects. Third, federal projects had this impact because they made agriculture feasible on land where it had previously been infeasible, and because, on average, they improved the quality of feasible land. Fourth, adjustment of land use to economic conditions was relatively gradual. On average, about 65 percent of forested acres which "should" have been converted to agriculture were converted in initial five-year periods; and about 38 percent of agricultural land which "should" have been abandoned were abandoned over initial five-year periods.

Simulations with estimated parameters show the quantitative importance of these effects. If there had been no federal flood-control or drainage projects constructed in the 36-county area after 1934, approximately 1.15 million fewer acres of forested wetlands would have been converted, 31 percent of total depletion. Long-term (steady state) depletion due to federal projects (constructed through the year 1984) is estimated to amount to more than 1.23 million acres, about 32 percent of estimated long-term depletion. Since the total acreage protected by projects was about 5.3 million acres, the results imply an average "propensity to convert" protected acres of about 0.22.

Of the factors considered in the econometric model, flood protection and drainage provision afforded by federal projects had the largest impact on net changes in forested acreage. The joint effect of natural topography and the mainline levee system was of secondary importance; and net forestry revenues and conversion costs exerted substantial restraints on wetland clearing.

In terms of public policy, the evidence highlights a striking inconsistency in the federal government's approach to wetlands. In articulated policies, laws, and regulations, the government recognizes the large positive externalities associated with wetlands; the Bush Administration has endorsed a "zero net-loss of wetlands" policy. But public investments in wetlands (in the form of flood-control and drainage projects) create major incentives to convert these areas to alternative uses. Clearly, the overall justification for these federal programs ought to be reexamined, and stringent tests of public need should be applied to both public and private actions which have direct or indirect effects on wetland resources.

The conclusion that major public infrastructure investments affect private land-use decisions (often thereby generating negative externalities) may not be a surprise to many readers of this journal, but the analysis described here provides evidence which contrasts sharply with the accepted wisdom among policymakers. It is hoped that the quantification of these effects will give these realities a more prominent role in policy debates. As wetlands, tropical rain forests, barrier islands, and other sensitive environmental areas become more scarce, their marginal social value rises. If induced land-use changes are not considered, we will engage in more and more public investment programs whose net social benefits are negative.

DATA APPENDIX

The data used in this study were collected as part of a previous research effort, sponsored by the U.S. Department of the Interior and carried out on behalf of the Environmental Defense Fund. For complete citations of sources of data and more information regarding the construction of requisite variable series, see Stavins 1986 and 1988. Also, a more comprehensive data appendix is available on request from the authors of the present paper.

Land-Use Patterns: The U.S. Forest Service periodically measures land use at sample sites, using a sampling procedure based upon aerial photographs. Sample locations are classified on a forest/non-forest basis, and these 0–1 observations are then converted into estimates of acreage of county land which is forested, S_{it}.

TABLE A1—SUMMARY STATISTICS FOR MAJOR VARIABLES[a]
1939 and 1984

Variable[b] Name	1939 Minimum	1939 Mean	1939 Maximum	1984 Minimum	1984 Mean	1984 Maximum
a	64.00	147.12	221.83	201.29	246.70	318.21
ac	62.41	155.38	197.51	117.92	180.97	295.46
ACCU	0.000	0.002	0.057	0.000	0.271	0.963
C	4.77	4.77	4.77	22.76	22.76	22.76
DRY	0.142	0.529	0.823	0.142	0.529	0.823
fn	−0.03	0.59	2.03	3.56	11.81	27.38
$\tilde{f}$	1.50	4.00	8.57	5.32	14.80	25.92
FORCH	−0.072	−0.020	0.080	−0.144	−0.023	0.016
PHDI	−2.150	−1.054	−0.020	−1.050	0.738	1.690
S	44.7	165.8	367.8	11.5	73.2	277.2
SCCU	0.000	0.000	0.000	0.000	0.100	0.864
T	129.9	360.4	590.7	110.9	354.6	590.7

[a] All monetary figures are real 1984 dollars.
[b] Variables are defined as follows:
a = annual agricultural revenue (dollars per acre);
ac = annual agricultural costs of production (dollars per acre);
ACCU = share of county protected by Army Corps of Engineers projects;
C = cost of conversion (dollars per acre);
DRY = share of county naturally protected from flooding;
fn = annuity of net forestry revenue minus windfall from clearcut;
$\tilde{f}$ = annuity of delayed net forestry revenue (dollars per acre);
FORCH = change in forestland as a share of total county over 5 years;
PHDI = Palmer hydrological drought index;
S = stock of forestland (thousand acres);
SCCU = share of county protected by USDA SCS projects; and
T = total county area (available for conversion, thousand acres.

Total county land available for conversion (not conserved by government) is represented by T_{it}. The net change in forest land as a share of all available county land is $FORCH_{it}$.

Agricultural Revenue: The average (gross) real agricultural revenue per acre, a_{it}, is a weighted average for four crops (soybeans, cotton, rice, and corn), based upon agricultural prices, production levels, yields, and acreages. Data on crop acreages, production levels, and prices come from the U.S. Census of Agriculture plus state publications. A weighted average of production costs, ac_{it}, was developed from state documents; costs considered were case expenses, which include variable plus fixed expenses (general farm overhead, taxes and insurance, and interest), but not capital replacement nor allocated returns to owned inputs.

Forestry Revenue: Annual forestry net revenue per acre, fn_{it}, consists of two components: the difference between the (annualized) revenue stream generated by periodic harvesting of limber and the (annualized) one-time revenue received from a clearcut of the forest prior to conversion. Thus, real forestry net revenue per acre is a weighted average of annual revenues from sawlogs and pulpwood minus the annuity of a windfall which is gained from a clearcut of timber if conversion is carried out. If farmland is abandoned and allowed to return to its forested state, there is a delay equal to the rotation length before harvests can commence. The annuity of delayed net forestry revenue is $\tilde{f}_{it}$ (see equation (8) in text).

Cost of Conversion: The time-series, C_t, is the average cost of conversion of wetlands to cropland. Because geographic variance in the cost of conversion is largely a function of soil moisture, a panel of conversion cost estimates were developed by allowing for the interaction of C_t and $PHDI_{it}$, as described above.

Artificial Flood Protection and Drainage Provision: Projects of the U.S. Army Corps of Engineers and the Soil Conservation Service were considered. In both cases, the primary measure of project impact was the "protected acreage" of projects, hydrologically defined as the area which experiences some reduction in the extent and frequency of flooding as a result of project construction. The respective Corps and SCS variables are $ACCU_{it}$ and $SCCU_{it}$, the sum of which is $PROJ_{it}$.

Natural Flood and Drainage Conditions: A measure of the average natural probability of flooding of sample counties, $FLRISK_i$, was developed from the National Resources Inventory, conducted by the Soil Conservation Service. The relevant variable for the analysis is the quantity, $DRY_i = 1 - FLRISK_i$.

Weather Conditions: The Palmer Hydrological Drought Index, estimated by the National Climatic Data Center, is related to precipitation, runoff, evapotranspiration, recharge, and soil water loss. Monthly drought index data were aggregated into quinquennial averages by county, $PHDI_{it}$.

REFERENCES

Arnott, Richard J. and Lewis, Frank D., "The Transition of Land to Urban Use," *Journal of Political Economy*, February 1979, 161–70.

Bohi, Douglas R. and Toman, Michael A., *Analyzing Nonrenewable Resource Supply*, Washington, DC: Resources for the Future, 1984.

Breusch, T. S., "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, December 1978, *17*, 334–55.

Brown, John P., *The Economic Effects of Floods: Investigations of a Stochastic Model of Rational Investment Behavior in the Face of Floods*, New York: Springer-Verlag, 1972.

David, Paul, "The Mechanization of Reaping in the Ante-Bellum Midwest," in Henry Rosovsky, ed., *Industrialization in Two Systems*, Cambridge: Harvard University Press, 1966, 3–39.

Durbin, J., "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors Are Lagged Dependent Variables," *Econometrica*, May 1970, *38*, 410–21.

Gardner, Bruce L. and Chavas, Jean-Paul, "Market Equilibrium with Random Production," Paper presented at AAEA Annual Meeting, Pullman, Washington, August 1979.

Gill, Richard D., "On Estimating Transition Intensities of a Markov Process with Aggregate Data of a Certain Type: 'Occurrences but No Exposures,'" *Scandinavian Journal of Statistics*, 1986, *13*, 113–34.

Godfrey, L. G., "Testing Against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables," *Econometrica*, November 1978, *46*, 1293–1302.

Gotz, Glenn A. and McCall, John J., *A Dynamic Retention Model for Air Force Officers*, Rand Report R-3028-AF, Santa Monica, CA, December 1984.

______, "Estimation in Sequential Decision-Making Models," *Economic Letters*, 1980, *6*, 131–36.

Griliches, Zvi, "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, October 1957, *25*, 501–22.

Kramer, Randall A. and Shabman, Leonard, A., *Development of Bottomland Hardward Tracts for Agricultural Use: The Influence of Public Policies and Programs*, Prepared for the U.S. Department of the Interior, Fish and Wildlife Service, Washington, DC, 1986.

Pakes, Ariel, "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, July 1986, *54*, 755–84.

Pomfret, Richard, "The Mechanization of Reaping in Nineteenth-Century Ontario: A Case Study of the Pace and Causes of the Diffusion of Embodied Technological Change," *Journal of Economic History*, June 1976, *36*, 399–411.

Pope, Rulon D., "Supply Response and the Dispersion of Price Expectations," *American Journal of Agricultural Economics*, February 1981, *63*, 161–63.

Rosenzweig, Mark R. and Wolpin, Kenneth I., "Heterogeneity, Intrafamily Distribution, and Child Health," *Journal of Human Resources*, 1988, *23*, 437–61.

Rust, John, "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, September 1987, *55*, 999–1033.

Shabman, Leonard, "Economic Incentives for Bottomland Conversion: The Role of Public Policy and Programs," in Kenneth Sabol, ed., *Transactions of the Forty-fifth North American Wildlife and Natural Resources Conference*, Washington, DC: Wildlife Management Institute, 1980, pp. 402–12.

Stavins, Robert N., "Conversion of Forested Wetlands to Agricultural Uses: An Econometric Analysis of the Impact of Federal Programs on Wetland Depletion in the Lower Mississippi Alluvial Plain," Final report by Environmental Defense Fund to U.S. Department of the Interior, May

1986.

———, "The Welfare Economics of Alternative Renewable Resource Strategies: Forested Wetlands and Agricultural Production," unpublished doctoral dissertation, Department of Economics, Harvard University, May 1988.

——— **and Jaffe, Adam B.,** *Forested Wetland Depletion in the United States: An Analysis of Unintended Consequences of Federal Policy and Programs*, Harvard Institute of Economic Research Discussion Paper No. 1391, Cambridge, MA: Harvard University, July 1988.

Theil, Henri, *Economic Forecasts and Policy*, Amsterdam: North-Holland, 1961.

White, Halbert, "A Heteroscedasticity-Consistent Covariance Matrix and a Direct Test for Heteroscedasticity," *Econometrica*, April 1980, *48*, 721–46.

Young, Douglas I., "Risk Preferences of Agricultural Producers: Their Use in Extension and Research," *American Journal of Agricultural Economics*, December 1979, *61*, 1063–70.

U.S. Army Corps of Engineers, *Tensas River Basin Excluding Bayou Macon, Louisiana, Project, Phase I—General Design Memorandum No. 25, Appendix 9, Land Use Analysis*, Vicksburg, MS: Vicksburg District Office, October 1981.

U.S. Department of the Interior, *The Impact of Federal Programs on Wetlands, Volume 1: The Lower Mississippi Alluvial Plain and the Prairie Pothole Region*, Washington, DC, October 1988.

Utility Functions That Depend on Health Status: Estimates and Economic Implications

By W. KIP VISCUSI AND WILLIAM N. EVANS*

Taylor's series and logarithmic estimates of health state-dependent utility functions both imply that job injuries reduce one's utility and marginal utility of income, thus rejecting the monetary loss equivalent formulation. Injury valuations have unitary income elasticity, and the valuation of non-incremental risk changes and effects of base risks follow economic predictions. (JEL 851,026,913)

In the basic von Neumann-Morgenstern framework, individual utility depends on a single attribute—one's wealth. In some contexts, the character of the lottery payoffs may be so sweeping that it transforms the utility function. Consider, for example, a utility function for a risk-averse individual so that utility increases with wealth, but at a diminishing rate. If one were to treat death as being tantamount to a drop in income, then one would obtain the unreasonable result that death boosts the marginal utility of income. This implausible result highlights the fallacy of treating death and other severe health effects as monetary equivalents.

Robert Eisner and Robert H. Strotz (1961) first noted this class of difficulties in their analysis of flight insurance. They suggested that a bequest function is a more appropriate formulation of the utility function after one's death. Modification of the standard utility theory approach to recognize the complications posed by other forms of state dependence has not posed any insurmountable difficulties, as the theory of state-dependent utility is now well developed.[1]

Economists have applied the state-dependent approach to a diverse set of economic problems involving irreplaceable effects, product safety, and accidents.[2] By far the most widespread use of this formulation is with respect to state-dependent variations with individual health status. Richard J. Zeckhauser (1970, 1973) and Kenneth J. Arrow (1974) developed analyses of health care and health insurance decisions in which the utility functions for good and ill health may assume quite different shapes. These formulations led to an overhaul of the economic analysis of the optimal structure of health insurance.

In particular, let there be n discrete states of the world indexed by $j = 1, \ldots, n$. Each state has an associated health level, health$_j$, and income level Y_j. One can then write individual expected utility EU as the sum of the utilities in each health state weighted by the probability of s_j that that health state$_j$ occurs, or

$$\mathrm{EU} = \sum_{j=1}^{n} s_j U(health_j, Y_j),$$

as in Charles E. Phelps (1973) and Arrow

*George G. Allen Professor of Economics, Department of Economics, Duke University, Durham, NC, 27706, and Assistant Professor of Economics, Department of Economics, University of Maryland, College Park, MD, 20742, respectively. The authors would like to thank Gregory M. Duncan, two anonymous referees, and seminar participants at various universities for helpful comments.

[1] See, for example, Arrow (1964), Jack Z. Hirschleifer (1970), Ralph L. Keeney and Howard Raiffa (1976), and Edi Karni (1985).

[2] See Zeckhauser (1970, 1973), Arrow (1974), Philip J. Cook and Daniel A. Graham (1977), Spence (1977), Viscusi (1978, 1979), Joseph Pliskin, Donald S. Shepard, and Milton C. Weinstein (1980), Daniel A. Graham (1981), George W. Torrance (1986), Steven M. Shavell (1987), Viscusi and Michael J. Moore (1987), and Victor R. Fuchs and Richard J. Zeckhauser (1987).

(1974). Ideally, one would like to obtain a continuous measure of health capital to analyze explicitly how individual health status alters the structure of utility functions and to explore the economic implications of health. Since our empirical analysis includes only two health states—good health and a post-injury state—a continuous measure of health cannot be constructed. Instead, we subsume the role of health into a state-dependent utility function as in Arrow (1974), or

$$\mathrm{EU} = \sum_{j=1}^{n} s_j U_j(Y_j).$$

More specifically, let the notation $U(Y)$ denote the utility function in good health and $V(Y)$ denote the utility function in the post-injury state. This differing notation denotes that there is a health state-dependent component of a stable preference relationship that we cannot estimate explicitly.

Two assumptions are pivotal. First, for any given level of income, one's overall level of utility is greater when in good health than in ill health, or $U(Y) > V(Y)$. This assumption is not controversial; nor does it distinguish the health state approach from earlier models that treated adverse health effects as being equivalent to financial losses. Second, to make any judgments about the extent of the optimal insurance one must make an assessment with respect to the influence of one's state on the marginal utility of income, for any income level. Optimal insurance coverage when there is actuarially fair insurance available will equate the marginal utility of income in each health state (for example, see Zeckhauser (1970, 1973), Arrow (1974), A. Michael Spence (1977), W. Kip Viscusi (1979)).[3] If ill health does not alter the marginal utility of income, for any given income level, then full insurance is optimal. If ill health lowers (raises) the marginal utility of income for any given income level, less (more) than full income insurance is desirable. Thus, the relative magnitudes of $U'(Y)$ and $V'(Y)$ are key empirical parameters. The assumptions one makes about the shape of the utility function govern the fundamental aspects of all of the economic results derived with such models.

In the extreme case of one's death, it is not controversial to assume that one's marginal utility declines after the adverse health effect. For other health outcomes, the justification for assuming a drop in marginal utility is less clearcut. There is no theoretical basis for determining the shape of the utility function in these instances.

In Section I we describe the set of data used to estimate state-dependent utility functions. We will use two empirical approaches. First, Section II imposes no functional form restrictions on the utility function other than a Taylor's series expansion. This unrestricted approach provides tests of the two key assumptions of the state-dependent approach—whether utility is greater in good health or ill health and whether the marginal utility of income is boosted by or reduced by adverse health outcomes. In Section III we impose a specific functional form on the utility function (a logarithmic utility function) and then estimate the utility function in each of the two health states. Section IV uses these results to address a variety of key, but previously unresolved issues, including: the income elasticity of the implicit value of an injury, the valuation of non-incremental changes in risk, changes in risk-dollar tradeoffs with a change in the base level of risk, and the optimal rate of replacement of worker earnings through workers' compensation insurance. Many of these findings are of substantial, independent economic interest.

I. Sample Description

Although there is a considerable literature on wage-risk tradeoffs, estimating individual utility functions is not feasible with standard sets of survey data. Figure 1 makes the source of the difficulty apparent. Let ABC be the frontier of offered wage-risk combinations available in the market. The worker selects the optimal job B from this frontier, where his locus of constant expected utility EU is tangent to the wage of opportunities frontier.

[3] Although the assumption of actuarial fairness is clearly unrealistic (i.e., it assumes the insurance provides a free service with no administrative costs), it is a frequent reference point in theoretical analyses.

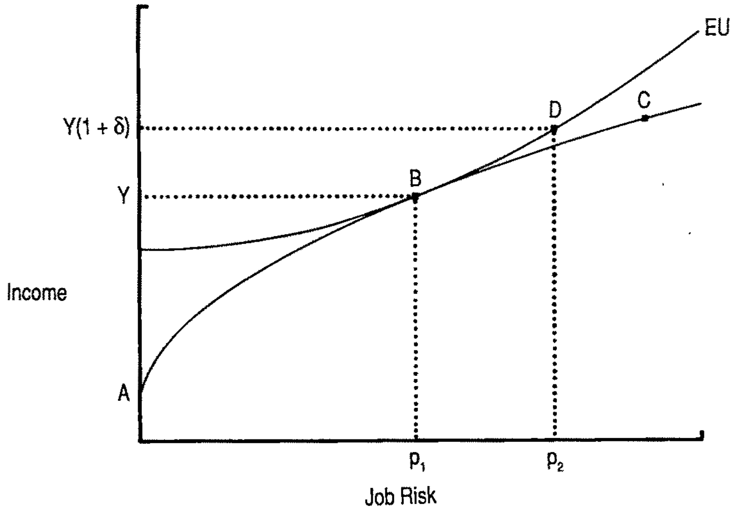


FIGURE 1. THE MARKET OFFER CURVE AND THE WORKER'S EXPECTED UTILITY LOCUS

Hedonic wage studies of compensating differentials for job risks involve estimation of the average rate of tradeoff for the equilibrium set of expected utility-wage offer tangencies observed in the market. The linear wage equation imposes a constant risk-dollar tradeoff, and a semilogarithmic formulation makes the tradeoff risk-dependent. Market data can only provide evidence regarding the slope of the observed tangencies with the frontier *ABC*. One cannot make any inferences regarding the shape of the individual worker utility functions except with respect to the rate of tradeoff at tangency with the opportunities locus.

We will follow an alternative approach of augmenting market data with reservation wage data obtained in response to different risk levels. In particular, we utilize the 1982 chemical worker survey by W. Kip Viscusi and Charles J. O'Connor (1984).[4] That analysis was primarily concerned with the economic implications of chemical labeling, whereas this paper is concerned with the utilization of the wage and risk information to estimate worker utility functions.

Table 1 summarizes the characteristics of the chemical worker sample. The personal characteristic variables included information on the worker's age (AGE), sex (MALE dummy variable-d.v.), marital status (MARRIED d.v.), number of children (KIDS), race (BLACK d.v.), years of experience at the firm (TENURE), and education (EDUC).

The survey also elicited the worker's pay for the period that was most meaningful for that particular occupational class. To promote comparability, these statistics have been converted to weekly, after-tax earnings, using average federal and state tax rates for the worker's income class and family status (i.e., marital status and number of dependents).[5] The workers in the sample averaged about $18,000 per year (1982 dollars) in after-tax income.

The key job attribute is the worker's perceived probability of an accident on his job, which is denoted by p_i, $i = 1,2$, where the

[4] More specifically, we will utilize the subsample of workers analyzed in Section III of Viscusi and O'Connor (1984). These workers experienced an increase in their job risk. Workers who were randomly assigned to the risk decrease experimental cell were not asked a reservation wage question.

[5] The tax adjustments were made using information provided in The Commerce Clearing House, *State Tax Handbook*, and *U.S. Master Tax Guide*.

TABLE 1—SAMPLE CHARACTERISTICS: MEANS AND STANDARD DEVIATIONS

Variable	Means (Std. Deviation)
AGE (in Years)	38.3 (11.8)
MALE (0 – 1 Sex Dummy Variable d.v.)	0.56 (0.50)
KIDS (Number of Children)	1.30 (1.53)
BLACK (0 – 1 Race d.v.)	0.06 (0.23)
TENURE (Years of Experience at Firm)	7.77 (6.97)
EDUC (Years of Education)	14.3 (3.37)
p_1 (Prior Probability of Accident)	0.084 (0.055)
p_2 (Posterior Probability of Accident)	0.249 (0.091)
Y (Weekly before Tax Earnings)	392.13 (161.52)
δ (Percent Wage Differential)	0.173 (0.150)
r_1 (Fraction Earnings Replacement by Workers' Compensation on Job 1)	0.637 (0.077)
r_2 (Fraction Earnings Replacement by Workers' Compensation on Job 2)	0.615 (0.093)
t_1 (Average Tax Rate on Job 1)	0.124 (0.047)
t_2 (Average Tax Rate of Job 2)	0.141 (0.052)
Sample Size	249

subscript 1 pertains to the pre-labeling situation and the subscript 2 pertains to the post-labeling situation. Workers assessed this probability using a linear risk scale on which there was indication of the average level of the risk for the entire private sector. Workers marked on the scale their assessed job risk with respect to this standardized injury scale, thus providing a risk metric scaled in terms of an annual job risk, that is, the assessed annual probabilities of injury are in the interval [0,1].

The risk metric is equivalent to the U.S. Bureau of Labor Statistics (BLS) private sector injury and illness rate. The risk assessments were 10 percent greater than the private sector average risk for that period and were 46 percent larger than the average chemical industry accident rate. Since BLS accident statistics do not capture the longer term health risks in the chemical industry, this pattern is quite reasonable.

The workers were randomly assigned to one of four different chemical labeling groups—asbestos, TNT, sodium bicarbonate, and chloroacetophenone (an industrial chemical that causes tearing). The workers were told that the chemical would replace the chemical with which they currently worked. Thus, there would be a change in the chemical used rather than a change in the labeling of the chemical with which the individual currently worked. Respondents then assessed the posterior risk, p_2, which is roughly triple the prior risk, Only workers who reported an increased risk assessment are included in the sample analyzed in this paper, since it was only for this group that reservation wage information was obtained. Almost all workers who reported a risk decrease were shown a label for sodium bicarbonate, which was the experimental treatment that corresponded to elimination of the chemical hazards.

The risk scale established a standardized reference lottery that the worker could use in assessing the risk equivalent of his job before and after seeing a warning label. Ideally, both the severity and duration of the reference injury should be the same, where the scale is used to establish differences in probabilities. Although this was the intent of the survey design, the significant differences in injury severity across the label treatment groups may have affected worker responses. In particular, the empirical properties of the responses by workers in the asbestos and TNT label groups are similar, but the behavior implied by the chloroacetophenone group is somewhat different. Since chloroacetophenone leads to only temporary eye irritation, and the other two chemicals pose risks of death, this difference is consistent with the character of the injuries. We will provide empirical estimates for each labeling subsample as well as for the full sample to explore differences across chemicals.

For workers who assessed their job risk as being greater, the survey ascertained the percentage wage increases δ needed to compensate the worker for the increased risk. The average δ value was just under 20 percent. Since the respondents were told that the results would be used for a doctoral dissertation at an institution in a different state and would not be disclosed to their employers, there was no apparent incentive for them to overstate their reservation wage. Moreover, the implicit value of a statistical injury reported in Viscusi and O'Connor (1984) and in Section IV of this paper are not excessively large and are in line with the literature.[6] Section III examines the effect on the estimates of potential response biases.

In terms of Figure 1, the survey first ascertained the information associated with point B—the base risk p_1 and the associated weekly earnings Y. It then altered the risk to a level assessed by the worker as being p_2, with an associated weekly earnings of $Y(1+\delta)$, which is point D. Thus, the survey includes information with respect to two points, B and D, on a constant expected utility locus EU where the expected utility EU_1 on the initial job equals the expected utility EU_2 after the wage and risk increase. The starting point and the post-labeling point D differ across workers so that in effect we observe 249 different pairs of points along 249 different utility functions. In contrast, the most that can be accomplished using observed market wage-risk data for this sample is to estimate ABC using the 249 points B. Since the survey generates only one equation, $\text{EU}_1 = \text{EU}_2$, we are not able to identify the shape of both U and V. However, we are able to generate relationships between the two utility functions, such as differences and ratios.

The final variable needed to complete the formulation of the payoffs in each state is the level of workers' compensation benefits after an injury. Since these benefits are not taxed, for comparability the income in the healthy state is in after-tax terms. The earnings replacement rate variable is based on the state benefit formulas for temporary total disability and the characteristics of the individual respondent. The benefit calculation takes into account the worker's income, benefit ceilings and floors, and the dependence of benefits on family characteristics.[7] The average replacement rate at the initial job (r_1) and for the experimental job (r_2) are both about two-thirds.

Let U denote the utility of wealth in good health, and let V denote the utility of money after a job injury. Then a wage increase that equates the expected utility that the worker obtains from his initial job and the transformed job satisfies.

$$(1)\quad (1-p_1)U(Y(1-t_1))+p_1V(Yr_1) = (1-p_2)U(Y(1+\delta)(1-t_2)) + p_2V(Y(1+\delta)r_2).$$

[6] The chemical worker sample yields rates of tradeoff that imply a value of \$10,000–\$20,000 per statistical injury reduced. If there were upward response bias, these estimates should exceed comparable values obtained using market data and hedonic wage equations. However, the estimates in the literature tend to be somewhat greater, as they cluster in the \$20,000–\$30,000 range. See the survey by Viscusi (1986) and the recent estimates by Viscusi and Moore (1987).

[7] The reasonableness of this approach to capturing empirically the role of workers' compensation is discussed in Viscusi and Moore (1987).

The worker reports his base earnings Y, his required wage increase δ, and his prior and posterior risk assessments, p_1 and p_2. Information regarding the worker's income level is used to construct the tax variables, t_1 and t_2, and the workers' compensation replacement rates, r_1 and r_2. The formulation in equation (1) takes into account the favorable tax treatment of workers' compensation benefits, as taxes only affect earnings in the good health state.

The principal empirical test of whether the state-dependent approach is valid is whether

$$U(Z) > V(Z) \tag{2}$$

and

$$U'(Z) > V'(Z), \tag{3}$$

for identical income levels Z in each state. Inequality (2) will be satisfied by both a health state and a monetary loss equivalent model since in each case ill health lowers welfare. The distinctive condition is defined by inequality (3). Under the state-dependent approach, inequality (3) could be in either direction, although the sign in inequality (3) is the more frequent assumption. With a monetary loss equivalent approach, an injury will boost the marginal utility of income for the usual risk-averse preferences, leading to a reversal of inequality (3). If inequality (3) is satisfied, we can reject the monetary equivalent model and the class of health in state models that do not alter the utility function in the manner indicated by inequalities (2) and (3).

II. Estimates with Unrestricted Functional Forms

Ideally, one would like to estimate equation (1) without imposing any restrictions on the shapes of U and V. However, with information on two particular points along the constant expected utility locus, one cannot estimate two different nonlinear functions with available data.

Two approaches are feasible. First, one can estimate specific features of the U and V functions without imposing functional form restrictions on their shape, as in this section. Second, one can impose constraints on the shapes that U and V can take, as in Section IV. These two different estimation approaches provide a robustness check on the results.

A. *First-Order Taylor's Series*

The procedure that we adopt in this section is to construct a first-order Taylor's series approximation of utility functions in each health state. The second-order Taylor's series terms, which we will explore in Section IIB, were not statistically significant.

For each of the utility functions, we will use the same level of income as the point of expansion, where this level is Y, the weekly before-tax income. From the definition of the Taylor's series, we generate the following approximations to utility:

$$\begin{aligned} U(Y(1-t_1)) &\cong U(Y) \\ &\quad + \{Y(1-t_1)-Y\}\,U'(Y), \end{aligned} \tag{4a}$$

$$V(Yr_1) \cong V(Y) + \{Yr_1 - Y\}\,V'(Y), \tag{4b}$$

$$\begin{aligned} U(Y(1+\delta)(1-t_2)) &\cong U(Y) \\ &\quad + \{Y(1+\delta)(1-t_2)-Y\}\,U'(Y), \end{aligned} \tag{4c}$$

and

$$\begin{aligned} V(Y(1+\delta)r_2) &= V(Y) \\ &\quad + (Y(1+\delta)r_2 - Y)V'(Y). \end{aligned} \tag{4d}$$

After substituting the values of (4a)–(4d) into equation (1), we obtain

$$\begin{aligned} &(p_2-p_1)(U(Y)-V(Y)) \\ &\quad = ((1-p_1)t_1-(1-p_2)(t_2+\delta t_2-\delta)) \\ &\quad\quad \times YU'(Y) + (p_1(1-r_1) \\ &\quad\quad - p_2(1-r_2-r_2\delta))YV'(Y). \end{aligned} \tag{5}$$

All of the variables in equation (5) are known except for those parameters involving the utility functions for each state. It is these terms that will be estimated.

Let

$$\text{(6a)} \qquad \beta_1 = U(Y) - V(Y),$$

$$\text{(6b)} \qquad \beta_2 = U'(Y),$$

and

$$\text{(6c)} \qquad \beta_3 = V'(Y).$$

The dependent variable will be the percentage wage compensation δ that the worker requires to face the increased risk. This empirical structure follows that of the questionnaire, as the survey asked workers how much additional compensation they required to work with the new chemical. Although workers could respond either in absolute or percentage terms, in each case it was the wage premium that was elicited. We can consider δ as the dependent variable since its value is conditioned on knowing all other variables —namely Y, p_1, and p_2.

Inserting the values for β_1, β_2, and β_3 from equation (6) into equation (5) and solving for the endogenous value δ yields

$$\text{(7)} \quad \delta = \left[\frac{(H_1\beta_2 + H_2\beta_3)Y - (p_2 - p_1)\beta_1}{\{(1-p_2)(t_2-1)\beta_2 - p_2 r_2 \beta_3\}Y} \right] + \varepsilon,$$

where

$$\text{(8a)} \quad H_1 = (1-p_1)t_1 - (1-p_2)t_2,$$

and

$$\text{(8b)} \quad H_2 = p_1(1-r_1) - p_2(1-r_2).$$

Introducing subscripts to denote the ith individual, we have thus hypothesized an empirical relationship of the form

$$\text{(9)} \qquad \delta_i = f(X_i, \beta) + \varepsilon_i,$$

where $f(\cdot)$ is a nonlinear function capturing the bracketed term on the right-hand side of equation (7), X_i is a $(k \times 1)$ vector of variables unique to the respondent (t_i, r_i, Y_i, etc.), β is a $(p \times 1)$ vector of parameters to be estimated, and ε_i is an i.i.d. error term. We can obtain an estimate of β via nonlinear least squares. The nonlinear least squares estimator[8] $\hat{\beta}$ is consistent so long as the error terms are independent and identically distributed, with mean zero and finite variance σ^2—assumptions that will be tested below.

Given the structure of equation (7) and the nature of the data, it is possible to estimate only two of the three parameters. With no loss of generality, set the coefficient

$$\text{(10)} \qquad \beta_2 = U'(Y) = 1.$$

The two tests that will be possible with the model are whether utility is greater in the healthy state, or

$$\text{(11)} \qquad \beta_1 = U(Y) - V(Y) > 0,$$

and whether ill health lowers the marginal utility of income, or

$$\text{(12)} \qquad \beta_3 = V'(Y) < 1.$$

A test of the financial loss model of adverse health effects would be to test the joint restriction $\beta_1 > 0$ and $\beta_3 > 1$. Thus, the distinguishing test of the health state approach, as compared with the financial loss model, is inequality (12).

Table 2 presents the nonlinear least squares estimate of equation (7), where β_2 has been constrained to equal 1. The first equation in Table 2 presents the estimates for the full sample, and the next three equations report estimates for each label subsample. The utility function parameters can be viewed as averages across the sample. The final equation allows each parameter to be a linear combination of the major human capital variables, providing evidence on heterogeneity of preferences. In this case we report both the individual parameter estimates as well as the estimated β_1 and β_3 values evaluated at the sample mean.

[8] A. Robert Gallant (1975) describes the nonlinear estimation procedure and its properties.

TABLE 2—NONLINEAR LEAST-SQUARES ESTIMATES OF FIRST-ORDER TAYLOR'S SERIES EXPANSION

	Coefficient Estimate, (Asymptotic Standard Error) and [Heteroscedastic Consistent Standard Error]				
Parameter	Full Sample	TNT	Asbestos	Chloroacetophenone	Full Sample
β_{10} (Intercept)					225.3 (70.92) [64.25]
β_{11} (EDUC)					−5.30 (5.28) [4.72]
β_{12} (TENURE)					1.22 (1.81) [1.98]
β_1	167.7 (11.81) [12.90]	194.1 (17.07) [17.36]	165.5 (15.63) [17.07]	−38.74 (60.31) [105.3]	158.9[a] (17.80) [19.17]
β_{30} (Intercept)					0.775 (0.826) [0.783]
β_{31} (EDUC)					−0.023 (0.043) [0.044]
β_{32} (TENURE)					−0.032 (0.018) [0.020]
β_3	0.773 (0.134) [0.163]	0.818 (0.194) [0.153]	0.415 (0.187) [0.173]	2.522 (0.561) [1.310]	0.859[a] (0.228) [0.225]
R^2	0.505	0.438	0.480	0.456	0.513
nR^2 Test[b]	1.469	1.123	2.166	20.19	11.20
Observations	249	78	87	84	249

[a] Evaluated at sample averages for EDUC and TENURE.

[b] This statistic for the first 4 columns is asymptotically distributed χ^2 with 3 degrees of freedom. The critical value for χ^2 (3 d.f.) at the 95 percent confidence level is 7.81. The statistic in the fifth column is asymptotically distributed as χ^2 (18 d.f.) with a 95 percent confidence level of 28.87.

The estimates of β_1 and β_3 are extremely precise and in the expected direction for all but the chloroacetophenone results. The coefficient β_1, which represents the difference between the utility when healthy and when injured, has the expected positive sign, with a coefficient that is over 10 times larger than its standard error for the first three sets of results. The β_1 coefficients for these first three columns also are not significantly different from each other. Since von Neumann-Morgenstern utility functions are defined only up to a positive linear transformation, it is only the sign of β_1 rather than its magnitude that is of consequence. Individuals prefer the good health state, as predicted.

The coefficient of β_3 is also positive and passes tests of statistical significance at very demanding levels since the asymptotic t-ratio is almost 6. The point estimate of β_3 for the full sample is 0.773, which implies that the marginal utility of income in the ill health state is about three-fourths that of the good health state. The confidence intervals of β_3 for TNT and asbestos overlap the confidence intervals of the full sample estimate of β_3.

The most pertinent statistical test is not whether β_3 is significantly different from zero but whether β_3 is significantly below 1, the marginal utility of income when healthy. One can reject the hypothesis that β_3 equals 1 at the 5 percent confidence level. The marginal utility of income is significantly lower in ill health than when in good health for the first three columns of estimates.

This result provides the distinguishing test between the financial loss and health state models of injuries. For all results except chloroacetophenone, there is evidence that the injury lowers welfare, so that β_1 will be positive. If this lowering takes the form of being tantamount to being a drop in income, then β_3 will exceed 1. With the health state model, the injury alters the shape of the utility function, with the most common assumption being that an injury reduces the marginal utility of income. The estimate of β_3 is in line with the health state model, and it suggests that treatment of health effects as being equivalent to monetary losses is inappropriate.

The one divergent set of results is for chloroacetophenone. The sign of β_1 is negative, which is the opposite of the expected relationship in inequality (11), but this coefficient is not statistically significant at the 5 percent level. The raised marginal utility of income after an accident ($\beta_3 = 2.522$) is consistent with the financial loss equivalent of the model rather than the health state approach. This result is not implausible since this chemical imposes no permanent health impairment. It is an eye irritant that causes tearing but does not inhibit one's ability to derive utility from additional expenditures. In Section III we will impose more structure on the estimation, which reduces but does not completely eliminate the differing performance of the chloroacetophenone group.

To explore possible heterogeneity in the parameter estimates, in the final equation in Table 2 the parameters vary across individuals according to the linear equation,

$$\beta_i = \beta_{i0} + \beta_{i1}EDUC + \beta_{i2}TENURE, \qquad \text{for } i = 1 \text{ and } 3,$$

where worker education (EDUC) and job experience (TENURE) are good measures of lifetime wealth.[9] Thus, for both β_1 and β_3 this equation includes estimates of this parameter and its interaction with education and job tenure. If better educated and experienced workers suffer a greater (lower) drop in utility after an injury, then the sign of the interaction with β_1 will be positive (negative). Similarly, a larger (smaller) drop in marginal utility after an accident will lead β_3 to have a negative (positive) sign.

Evaluating the sum of the parameter estimates at the sample averages, we find $\beta_1 = 158.9$ and $\beta_3 = 0.859$. These results are within one standard error of the full sample estimates in column 1 of Table 2, even when the adjusted standard errors are used.[10] Because of the large variances for the variables in the final equation, the estimated β_3 is less than one standard deviation away from the critical value of 1 that is pertinent for the testing of the hypothesis given in inequality (12) above. None of the personal characteristic variables is statistically significant at the 5 percent level (two-tailed test). The most precisely estimated interaction is the β_3 interaction with tenure (significant at the 5 percent level, one-tailed test), which suggests that more experienced workers will suffer a greater drop in marginal utility after an accident. This result is expected since the greater family responsibilities and more limited mobility of more senior workers creates a demand for greater insurance coverage after an accident, which is the substantive implication of a lower value of β_3.

One extreme hypothesis that can be tested using the results in Table 2 is whether injuries have no effect whatsoever on either the level of utility or the marginal utility of income. This hypothesis can be rejected at even very demanding confidence levels in every case shown in Table 2.

Two statistical issues must be addressed before turning to alternative specifications of

[9] More detailed sets of interaction created convergence problems and could not be estimated.

[10] The adjusted standard errors are developed using the procedure in White and Domowitz (1984). This procedure is designed to adjust for the influence of heteroscedasticity.

the utility function. First, the i.i.d. assumption of the model may not be satisfied for this cross-sectional data base, as the error term may be heteroscedastic. Table 2 includes both the conventional standard error and the heteroscedastically consistent standard error for nonlinear models.[11] The second set of standard errors has very similar indications for statistical significance so that any heteroscedasticity that is present does not appear to be consequential.

Second, it is possible to test the i.i.d. assumption explicitly and to test the correctness of the model specification. Based on the Halbert L. White and Ian Domowitz (1984) test summarized in the Appendix, one cannot reject at the 95 percent confidence level the assumption that the errors are homoscedastic. Furthermore, one cannot reject the assumption that the first-order Taylor's series model for the full sample is a correct specification up to an independent additive error term.

B. *Second-Order Taylor's Series*

To test the robustness of the previous model and to explore the potential role of second-order terms, we also estimated a model based on a second-order Taylor's series. The second-order expansion will allow us to estimate $U''[\cdot]$ and $V''[\cdot]$, which can be used to calculate measures of risk aversion. The second-order expansion is substantially more difficult to solve algebraically because the expansion of $U(Y(1+\delta)(1-t_2))$ and $V(Y(1+\delta)r_2)$ about Y generate a quadratic expression in δ, our variable of interest. The second-order model can be constructed as follows. Denote the quadratic expression in δ as

$$A\delta^2+B\delta+C=0,$$

and define β_1, β_2 and β_3 as before. Let $\beta_{22}=U''(Y)$ and $\beta_{33}=V''(Y)$.

Some straightforward (but lengthy) algebra generates the following values for the quadratic coefficients:

$$\mathrm{EU}_1=\beta_1+(1-p_1)\{-\beta_2Yt_1+0.5(Yt_1)^2\beta_{22}\}$$
$$+p\{Y(r_1-1)\beta_3+0.5[Y(r_1-1)]^2\beta_{33}\},$$
$$A=0.5(1-p_2)\{[Y(1-t_2)]^2\beta_{22}$$
$$+p_2(Yr_2)^2\beta_{33},$$
$$B=(1-p_2)Y(1-t_2)(\beta_2-Yt_2\beta_{22})$$
$$+p_2Yr_2\{\beta_3+Y(r_2-1)\beta_{33}\},$$

and

$$C=\beta_1+(1-p_2)\{-Yt_2\beta_2+0.5(Yt_2)^2\beta_{22}\}$$
$$+p_2\{Y(r_2-)\beta_3+0.5(Y(r_2-1))^2\beta_{33}\}$$
$$-\mathrm{EU}_1.$$

The solution to the quadratic suggests two possible roots. In preliminary analysis with this expansion, numeric calculations indicate that only one of the roots can predict positive values for δ, and therefore, the implicit equation we choose to estimate is of the form:

$$\delta=\left[-B+(B^2-4AC)^{1/2}\right]/2A+\varepsilon.$$

As is the case with the first-order series, both β_2 and β_3 (the marginal utility terms) are not identified and so without loss of generality, we set $\beta_2=1$ and test whether $\beta_3<1$. Consumer theory suggests that both second-order terms should be negative, but there is no theoretical basis for predicting which term should be larger in absolute value.

In the first column of Table 3, we present Taylor's series results for a model where both of the second-order terms are allowed to vary. The value of β_3 drops substantially from the 0.77 estimated in the first-order case. However, we accept the hypothesis that the coefficient is significantly below the critical value of 1 at about the same confidence level as in Table 3, which is the main hypothesis of interest. The magnitude and the

[11] *Ibid.*

TABLE 3—NONLINEAR LEAST SQUARES ESTIMATES OF SECOND-ORDER TAYLOR'S SERIES EXPANSION

Parameter	Coefficient Estimate, (Asymptotic Standard Error) and [Heteroscedastic Consistent Standard Error] Full Sample	Full Sample $\beta_{22}=\beta_{33}$
β_1	184.2 (13.47) [15.63]	170.3 (11.79) [12.33]
β_3	0.261 (0.316) [0.350]	0.701 (0.188) [0.189]
β_{22}	2.1E−3 (2.4E−3) [2.4E−3]	−7.4E−4 (1.5E−3) [1.1E−3]
β_{33}	−1.8E−3 (1.6E−3) [1.7E−3]	−7.4E−4 (1.5E−3) [1.5E−3]
R^2	0.509	0.506
nR^2 Test	33.86[a]	13.89[b]

[a] This statistic is asymptotically distributed χ^2 with 10 degrees of freedom. The critical value for χ^2 (10 d.f.) at the 95 percent confidence level is 18.31.

[b] This statistic is asymptotically distributed χ^2 with 6 degrees of freedom. The critical value for χ^2 (6 d.f.) at the 95 percent confidence level is 12.59.

significance of β_1 is quite similar to the first-order result. The coefficients for the second-order terms are not estimated with a great deal of precision, and we cannot reject the joint hypothesis that both terms equal zero.

Given the lack of precision in the second-order terms, we estimated a second model reported in the final column of Table 3, where the second-order terms are restricted to be equal. By restricting $\beta_{22}=\beta_{33}$, the estimates for the second-order terms are both negative but insignificant, and the estimates for β_1 and β_3 are quite similar to the first-order series results. Although the estimate for β_3 is quite different in both columns in Table 3, we cannot reject the hypothesis that the results in both columns are equal.[12] This is not surprising given the closeness of the estimates for β_1 in both columns, the imprecision in both second-order terms, and the large variance for β_3 in column 1.[13]

Although the second-order results are consistent with the theoretical predictions, the failure of any second-order terms to be statistically significant suggests that the earlier first-order results represent a reasonable approximation. The specification test results reinforce this conclusion.

III. Estimates with Logarithmic Utility Functions

To obtain estimates of the entire utility function shape, as opposed to simply the

[12] The test statistic is asymptotically an F with 1 and 245 degrees of freedom. The test statistic was 1.81, which is below the 95 percent critical value of 3.84.

[13] Notice also that the nR^2 test indicates possible model misspecification. The large variances for the second-order terms and the rejection of the nR^2 test are not surprising given that the survey asked for only one response. The survey was originally designed to elicit the response ΔY for a given Δp. Without placing more structure on the utility function, it may be difficult to determine second moments of utility since respondents were not asked additional questions to indicate changes in the rate of tradeoff between p and Y.

utility differences and the relative marginal utility of the two states, one must impose additional structure on the utility functions. The particular functional form we have selected is the Cobb-Douglas parameterization, where $U(Y)$ is of the form Y^u. Although this functional form does not have the same flexibility as the approximation involving the Taylor's series expansion, it is not extremely restrictive.[14] Upon taking logarithms of the within-state utility, we obtain a logarithmic utility function where $U(Y)$ equals $u[\log(Y)]$.

In the case of this model, the logarithmic formulation implies that

$$U(Y)=u[\log(Y)], \tag{13}$$

and

$$V(Y)=v[\log(Y)], \tag{14}$$

where u is a multiplicative parameter for the healthy state utility function and v is the parameter for the unhealthy state. If $u>v$, then the utility and the marginal utility of income are greater when the worker is in the good health state, which is the standard assumption in the literature.

The logarithmic utility function is frequently used in finance contexts[15] and in empirical applications analyzing von Neumann-Morgenstern utility functions. This function embodies decreasing risk aversion, which is a common empirical phenomenon.[16]

The major drawback of the logarithmic approach is that the utility and marginal utility will each be governed by a single parameter. This formulation in effect links the tests of whether there is a utility drop in the ill health state (i.e., $v<u$) and whether there is a marginal utility decline (i.e., $v/u<1$), so that it does not provide an unconstrained test of the model. However, the overall test of behavior was the subject of the Taylor's series test, and one can view the logarithmic utility function as imposing more specific functional structure on the relationships that were shown to hold in Section II. The purpose of the additional structure is to obtain estimates that will be used in greater detail to examine attitudes toward risk. Although other functional forms for utility functions have appeared in the literature, these could not be used because of both the nature of the data and the iterative search procedure that was used.[17]

There are several types of checks on the realism of the model. Many of these checks are based on comparison with the Taylor's series estimates. First, does utility drop in the ill health state? The magnitude of any such drop is irrelevant since von Neumann-Morgenstern utility functions are unaffected by a positive linear transformation. Second, does the point estimate of the ratio of the marginal utility in ill health relative to good

[14] Melvyn A. Fuss, Daniel McFadden, and Yair Mundlak (1978) discuss its use in production contexts.

[15] Examples of the use of logarithmic utility functions in the finance literature abound. See Albert L. Kraus and Robert H. Litzenberger (1975), Mark E. Rubinstein (1977), and Paul A. Samuelson (1969).

[16] See Keeney and Raiffa (1976).

[17] We also attempted to estimate equation (1) with two other specifications for utility: the constant risk aversion (CRA) utility function and the constant relative risk aversion (CRRA) functions. However, these models are not identified given the formulation of the problem, as presented in equation (1). The CRA utility function is typically denoted as $U=-\exp[-rY]$, where $r<0$. given the CRA specification, we are unable to obtain a closed-form solution for δ. Instead, we attempted to estimate the implicit equation $\mathrm{EU}_1-\mathrm{EU}_2=\varepsilon$, where ε is an i.i.d. error with mean zero. Since there is no appropriate normalization of the parameters, we must estimate two variables, r_1 in a healthy state, and r_2 in an unhealthy state. Given the properties of the CRA function, the sum of squared errors is minimized where $r_1=r_2=0$, which forces utility to equal one in all periods. Subsequently, the parameter values generate the equality

$$\mathrm{EU}_1-\mathrm{EU}_2=[(1-p)+p]-[(1-q)+q]=0.$$

The CRRA utility function is defined to be $U=Y^{(1-\rho)}/(1-\rho)$, where $\rho\geq 1$. As in the previous example, we are unable to obtain a closed form solution for δ so we must estimate the implicit equation $\mathrm{EU}_1-\mathrm{EU}_2=\varepsilon$. Algorithmically, the sum of squared errors can be minimized by choosing extremely large values for both ρ_1 and ρ_2. This has the property of forcing utility in all periods to machine zero and therefore the difference, $\mathrm{EU}_1-\mathrm{EU}_2$, is also zero.

health differ statistically in the Taylor's series and logarithmic utility function cases? Third, do the models provide similar estimates of $[U(Y)-V(Y)]/V'(Y)$, which is a utility difference statistic that is unaffected by positive linear transformation of the utility function? Finally, we will provide a White-Domowitz (1984) test of whether the model is correct up to an additive independent error, thus providing a formal specification test.

The requirement given by equation (5) above can be written as

$$(15)\quad (1-p_1)u\{\log[Y(1-t_1)]\} + p_1 v\{\log(Yr_1)\} = (1-p_2)u\{\log[Y(1+\delta)(1-t_2)]\} + p_2 v\{\log[Y(1+\delta)r_2]\},$$

which equates the expected utility of the initial job and the transformed job.

Even in conjunction with the imposed functional form, it will not be possible to estimate both parameters, u and v. As a result, we will estimate their ratio α given by

$$(16)\qquad \alpha = u/v.$$

As in the Taylor's series case, the dependent variable is δ, the percentage wage increase that the respondent requires to face an increased risk. Using the normalization of equation (16), we solve for δ to yield

$$(17)\qquad \delta = \exp\left[\frac{K_1 - K_2}{(1-p_2)\alpha + p_2}\right] - 1 + \varepsilon,$$

where

$$K_1 = (1-p_1)\alpha\log[Y(1-t_1)] + p_1\log[Yr_1],$$

and

$$K_2 = (1-p_2)\alpha\log[Y(1-t_2)] + p_2\log[Yr_2].$$

We will estimate equation (17) using nonlinear least squares, again assuming the error term is i.i.d. with mean zero and finite variance.

The principal hypothesis is that

$$(18)\qquad \alpha > 1,$$

or, for any given level of income, both the level of utility and the marginal utility of income are higher in the good health state.

Two cases will be considered. First, we will estimate the homogeneous preference model in which all utility functions are identical (i.e., $\alpha = \alpha_0$), thus providing an average value for α across the sample. Second, we then permit α to vary with personal characteristic variables X_i. Doing so leads to the heterogeneous preference assumption that

$$(19)\quad \alpha = \alpha_0 + \beta_1 EDUC + \beta_2 TENURE,$$

where the personal characteristic variables are education and job experience.

The estimate of the homogeneous preference model appears as equation (1) in Table 4. The estimate of α_0 is clearly statistically different from zero (asymptotic $t=134$), but the more relevant issue is whether α_0 differs from 1.0. The estimate for the full sample that α_0 equals 1.077 lies 9.6 standard deviations above 1.0, so one can reject the hypothesis that $u = v$, indicating that both the utility level and the marginal utility are greater in the good health state.

The α_0 values for each of the chemical subsamples are also above 1.0. Both the TNT and asbestos α_0 values are not significantly different from each other or the full sample results. However, one can reject the hypothesis that all of the α_0 coefficients are identical. Nevertheless, the magnitude of the α_0 value for chloroacetophenone is not greatly different in magnitude (for example, its α_0 confidence interval overlaps with that for asbestos), and the overall structure of the utility function implied by the results is very similar to that for the other chemical subsamples. The additional structure imposed by the logarithmic utility function may have muted some of the differences across label-

TABLE 4—NONLINEAR LEAST-SQUARES ESTIMATES OF LOGARITHMIC UTILITY MODEL

	Coefficient Estimate, (Standard Error), and [Heterscedastic Consistent Standard Error]				
	Model 1				Model 2
Parameter	Full Sample	TNT	Asbestos	Chloroacetophenone	Full Sample
Intercept					1.294 (0.033) [0.033)
EDUC					−0.013 (0.002) [0.002]
TENURE					−0.004 (0.001) [0.001]
α_0	1.077 (0.008) [0.009]	1.094 (0.013) [0.014]	1.065 (0.012) [0.015]	1.043 (0.018) [0.022]	1.082[a] (0.016) [0.016]
R^2	0.363	0.147	0.208	0.436	0.466
nR^2 Test of Specification	12.00[b]	1.63[b]	6.55[b]	1.03[b]	3.13[c]
Observation	249	78	87	84	249

[a] Evaluated at sample averages for EDUC and TENURE.

[b] This test statistic is asymptotically distributed χ^2 with 1 degree of freedom. The critical value for χ^2 (1 d.f.) at the 95 percent confidence level is 3.84.

[c] This statistic is asymptotically distributed χ^2 with 6 degrees of freedom. The critical value for χ^2 (6 d.f.) at the 95 percent confidence level is 12.59.

ing groups that were apparent in the Taylor's series results.

The relative discrepancy between the marginal utility in the two health states is narrower for the logarithmic model than the first-order Taylor's series expansion. The ratio of marginal utilities, $V'(Y)/U'(Y)$, is given by $1/\alpha$ for the logarithmic model and by β_3 in the case of the Taylor's series model. The estimates are 0.93 and 0.78, respectively. In each case, the accident lowers the utility and marginal utility of income, but the Taylor's series estimates imply a greater relative gap in the marginal utilities. The 95 percent confidence intervals for $V'(Y)/U'(Y)$ for the two different estimation approaches overlap (full sample results).

Another comparison that is meaningful, given possible differences in the utility metric, is the ratio of the utility difference to the marginal utility of income when injured. One establishes a comparable metric for utility differences by dividing by a marginal utility term. The value of $[U(Y)-V(Y)]/V'(Y)$ is 216 for the full sample Taylor's series results and is 179 for the full sample logarithmic results—a difference of under 20 percent.

As in the case of the Taylor's series results, the adjusted standard errors are similar to those that have not been adjusted for heteroscedasticity. In addition, for Model 2 (but not Model 1), one cannot reject the hypothesis that the specification is correct up to an additive independent error (see Appendix). This result is not inconsistent with a similar finding for the first-order Taylor's series model since one can view the flexible form of the Taylor's series model as providing an approximation to the logarithmic formulation.

TABLE 5—SIMULATION RESULTS, SENSITIVITY OF ESTIMATES TO OVERESTIMATE OF δ

Percent Reduction in δ	Coefficient Estimates and Standard Errors				
	Taylor's Series Model			Logarithmic Model	
in percent	β_1	β_2	R^2	α	R^2
5	158.2	0.740	0.513	1.069	0.367
	(11.25)	(0.125)		(0.007)	
10	147.9	0.707	0.522	1.061	0.377
	(10.67)	(0.116)		(0.007)	
25	118.2	0.608	0.544	1.036	0.377
	(8.89)	(0.091)		(0.006)	

The imposition of additional structure with the logarithmic model has also facilitated the estimation of variations in individual preferences. Recall from equation (19) that positive coefficients imply that the particular demographic group has a higher utility of income value and higher marginal utility of income when in good health relative to ill health. Both education and tenure have negative signs, implying that there is less of a drop in the marginal utility after an accident for these workers.

A possible bias in the model may arise if workers responded strategically to the survey, for example, by exaggerating claims of the required wage increase δ. To check our results against a possible bias in δ, we reestimate the Taylor's series and the logarithmic model after systematically depressing the response variable δ. This simple test illustrates whether our results are sensitive to exaggerated claims of δ. Table 5 reports the test results from reducing the values of δ by 5, 10, and 25 percent.

The sensitivity analysis for the logarithmic model indicates that the parameter α is sensitive to a possible bias in the response variable δ. However, even a 25 percent overestimate in δ does not alter the basic conclusion that α is significantly greater than 1. Likewise, the general character of the results for the Taylor's series case are not altered as δ is decreased. As the percentage reduction in δ is increased, the primary parameter of interest, β_3, actually declines in value, indicating that the choice between the health state and the monetary loss model is not biased by a strategic response for δ.

IV. Economic Implications

Without knowledge of the shape of individual preferences, the domain of economic inquiry is largely limited to a single issue—the local rate of tradeoff between risk and money. Using the estimates of the logarithmic utility function, we will extend the domain of inquiry to assess how risk-money tradeoffs vary with the base level of risk, the extent of the risk change, and individual income. We will also estimate the optimal workers' compensation replacement rate. Knowledge of the utility function enables us to address a variety of concerns that have been central to the risk bearing field but which have never been addressed empirically.

A. *Variation in the Implicit Value of Statistical Injury with the Base Risk*

The most useful means for expressing the risk-money tradeoff is in terms of the dollar compensation required per unit of risk. This rate of tradeoff can be calculated for marginal changes of risk as it represents the value of $\partial Y/\partial p$ for a given value of expected utility.[18] At the mean risk level for the

[18] Using the formula in Viscusi (1979, p. 12), the expected utility formulation from the left side of equation (15), and assuming that an individual works 50

TABLE 6—EFFECT OF THE BASE RISK LEVEL ON THE IMPLICIT VALUE OF AN INJURY

	Implicit Dollar Value of Injury	
Base Risk Level	Logarithmic	Taylor's Series
0.0	13,262	11,313
0.1	13,357	11,569
0.2	13,454	11,838
0.3	13,553	12,119
0.4	13,653	12,414
0.5	13,754	12,724
0.6	13,857	13,050
0.7	13,961	13,392
0.8	14,067	13,753
0.9	14,174	14,134
1.0	14,284	14,537
0.085 (Sample Mean)	13,343	11,530

sample of approximately 0.085, the logarithmic utility function estimates yield a value of an injury of $13,343 (1982 dollars). As the value of injury figures reported in Table 6 indicate, there is not substantial variation in the implicit value of an injury with the base risk level, as the range is from $13,262 to $14,284.

The Taylor's series results can be used to generate similar estimates of the implicit value of injury, which is $\partial Y/\partial p$, holding EU_1 constant.[19] The implicit value of injury for the unconstrained second-order Taylor's series estimates is $11,530, which is somewhat below the logarithmic estimate. As the results in the final column of Table 6 indicate, the variation in the implicit value of an injury with the risk level follows the same general pattern as in the logarithmic case, but is somewhat greater. Using the estimates in which we constrain the statistically insignificant second-order terms to equal zero (see final column in Table 3), the implicit value of an injury rises to $12,057. This estimate is closer to the logarithmic utility function result.

The variation of the injury value with the risk level is of independent economic interest. Several economic models predict that the valuation of a risk change should be an increasing function of the base risk level.[20] The source of this effect is the lower opportunity cost of resources with high risk levels. At high levels of risk, the probability of spending the money when in good health is less. Since the utility and marginal utility of money is less when one is injured than when one is healthy, for any given income level, the additional expected utility produced by wage compensation is reduced by increases in the base risk.

The types of variations that are predicted theoretically are borne out by the results in Table 6 for both the logarithmic and Taylor's series cases. The additional compensation required to accept an increase in risk is greater for high base risks. The Table 6 results also indicate a change in the tradeoff with the base risk. These patterns follow economic predictions, as $\partial Z/\partial p > 0$ and $\partial^2 Z/\partial p^2 > 0$.

B. *Income Elasticity of the Value of an Injury*

On a theoretical basis, the value of an injury should increase with individual income and wealth, and available labor market

weeks per year, the implicit value of a statistical injury Z in the logarithmic utility function case is given by

$$Z = 50Y\frac{[\alpha \ln(Y(1-t)) - \ln Yr]}{(1-p)\alpha + p}.$$

[19] The statistic can be written as

$$Z \equiv \frac{\partial Y}{\partial p} = \frac{U(Y(1-t)) - V(Yr)}{(1-p)U'(Y(1-t)) + pV'(Yr)}.$$

The explicit characterization of Z in the Taylor's series model is generated by using a first-order series expansion to approximate $U[Y(1-t)]$ and $V[Yr]$ about Y, and a first-order series to approximate $U'[Y(1-t)]$ and $V'[Yr]$ about Y. Defining the parameters β_1, β_2, β_3, β_{22}, and β_{33} as before, and assuming the worker is employed for 50 weeks, we can write Z as

$$Z = 50\frac{\beta_1 - Yt\beta_2 - Y(r-1)\beta_3}{(1-p)\beta_2 + p\beta_3 - (1-p)Yt\beta_{22} + pY(r-1)\beta_{33}}.$$

The calculation uses the Taylor's series results from the second column of Table 3 and the mean risk level for p.

[20] For discussion of this and related issues, see Viscusi (1979) and Weinstein, Shepard, and Pliskin (1980).

data are consistent with this relationship (see Viscusi (1978, 1979)). However, the extent of the observed income effects have not been large since existing data sets are not well-suited to disentangling the role of compensating differentials for risk and income effects that govern job choice.

At the mean value of an injury for the sample, the income elasticity of the value of an injury in the logarithmic case is 1.0995.[21] Thus, the value of an injury is roughly proportional to one's base income level. Using Taylor's series results from the final column in Table 3, the parameter estimates suggest that the elasticity in the Taylor's series case is approximately 0.67.[22] In the health insurance context, estimated income elasticities are generally lower—typically 0.5 or less.[23] One might expect the health insurance income elasticity to be below the elasticity of the value of an injury since demand will be muted to the extent that additional health expenditures have a diminishing, probabilistic effect on one's well-being.

The estimates of the income elasticity of injuries indicate that the value placed on individual health is not a constant, but exhibits substantial heterogeneity (see Viscusi, 1979, 1983). Knowledge of the income elasticity of the value of statistical injuries is likely to be particularly useful in the valuation of government programs with long-term effects since the growth in income over time will boost the value of the risks reduced, offsetting much of the influence of discounting.

C. *The Value of Non-Incremental Risk Changes*

In some cases the risk change that must be valued does not involve a small incremental change in the probability. Although medical contexts create the greatest opportunities for quantum changes in the risk level, changes in large individual risks resulting from government regulation (for example, seatbelt use requirements) pose similar problems.

From an economic standpoint, individuals should exhibit a diminishing marginal valuation of risk reduction and an increasing marginal acceptance price for risk increases. Market risk data do not enable one to address these issues since the observed risk changes tend to be small.[24]

Knowledge of utility functions enables one to make such assessments, as Table 7 summarizes the value of non-incremental risk changes from the starting point of the mean injury risk of 0.085. The purchase of a risk reduction of −0.085 is tantamount to complete elimination of the risk. Using the logarithmic estimates, there is an associated value per unit risk reduction of $12,865 for such a complete elimination of the risk. Similarly, there is a $8,989 value for the first-order Taylor's series estimates. At the opposite

[21] For the logarithmic case, the income elasticity ε of the value of an injury is given by

$$\varepsilon = \frac{Y}{Z}\frac{\partial Z}{\partial Y} = 1 + \frac{\alpha - 1}{\alpha \ln(Y(1-t)) - \ln(Yr)},$$

which is obtained by using the value of Z from fn. 18.

[22] We can calculate ε for the second-order Taylor's series model by differentiating Z with respect to Y and multiplying by Y/Z. Assuming $U'''(\cdot) = V'''(\cdot) = 0$, the value for ε can be written as

$$\varepsilon = Y\frac{\beta_2 - \beta_3 - tY\beta_{22} - Y(r-1)\beta_{33}}{\beta_1 - tY\beta_2 - Y(r-1)\beta_3}$$

$$- \frac{Y}{(1-p)\beta_2 + p\beta_3 - (1-p)tY\beta_{22} + Y(r-1)\beta_{33}}.$$

Given the imprecision with which the second-order terms are estimated, we use the results from column 2 of Table 3 where β_{22} is restricted to equal β_{11} to calculate the value for Z. Income elasticities cannot be derived using the first-order results.

[23] The income elasticity for health insurance estimates and the underlying theory are discussed in Joseph P. Phelps (1973), Charles E. Newhouse and Phelps (1976), and Phelps (1987).

[24] The predicted pattern of behavior has been borne out in an experimental consumer context by W. Kip Viscusi, Wesley A. Magat, and Joel C. Huber (1987). In their study, the marginal valuations of successive risk reductions were elicited directly, whereas here we will estimate the value of non-incremental changes using the estimated utility function.

TABLE 7—DEPENDENCE OF THE IMPLICIT VALUE OF AN INJURY ON THE EXTENT OF THE RISK CHANGE

Risk Increment from a Sample Mean (0.085)	Implicit Dollar Value of an Injury	
	Logarithmic	Taylor's Series
−0.085	12,865	8,989
−0.050	13,059	9,131
−0.010	13,286	9,299
+0.010	13,401	9,386
+0.050	13,637	9,563
+0.250	14,918	10,566
+0.500	16,792	12,157
+0.750	19,041	14,314
+0.915	20,777	16,213

extreme, workers would require a risk-dollar tradeoff of \$20,777 (logarithmic) or \$16,213 (Taylor's series) to incur an increase in the injury probability from 0.085 to 1.0, or a risk increase of +0.915.

These results suggest how non-incremental risk changes differ in value from estimates based on small marginal changes in the risk. The value of an injury of \$13,343 (logarithmic) for incremental changes at the sample mean is only 4 percent larger than the injury value associated with complete elimination of the risk since the initial injury probability is close to zero. The injury value associated with risk increases to a risk of 1.0 is 56 percent greater (logarithmic) than the value at the mean since the risk change is quite substantial. In addition, the change in the value of an injury increases at an increasing rate as the risk level rises. For example, the implicit value of an injury (logarithmic) rises at 1.67 times the rate over the interval (+0.750, +0.915), as compared with the interval that it did over (0.0, +0.250). A similar pattern is observed in the Taylor's series case. Individuals demand increasingly large prices per unit risk for successive risk increases and are willing to pay successively smaller amounts for additional risk decreases, as predicted.[25]

D. *The Optimal Workers' Compensation Earnings Replacement Rate*

At present, the workers' compensation earnings replacement rate is based on an algorithm that typically set the benefit equal to two-thirds of the worker's gross wage rate, subject to certain minimum benefit levels, maximum benefit levels, and benefit duration amounts. Since the marginal utility of income is reduced by an injury, as our results for both the Taylor's expansion and logarithmic model indicate, less than full earnings replacement is desirable.

How much earnings replacement is desirable cannot be determined based on available labor market data.[26] Using worker utility functions, a precise assessment is possible. In particular, suppose that workers must purchase workers' compensation through an insurance market. If the risk of injury is p, then the price of actuarially fair insurance is $p/(1-p)$ and the cost of buy-

[25] See Viscusi, Magat, and Huber (1987) for a derivation and review of the antecedents in the literature. Milton C. Weinstein, Donald S. Shepard, and Joseph Pliskin (1980) present a related discussion for fatality risks.

[26] The most that can be done is to assess the wage offset that workers are willing to accept in return for workers' compensation benefits and compare this offset with what would be observed if insurance were optimal. The estimates in Viscusi and Moore (1987) imply that the levels of benefits were suboptimal in the 1970s, but the extent of the suboptimality could not be determined. Estimates for the 1980s in Moore and Viscusi (forthcoming) indicate that substantial increases in the benefit levels since the 1970s have led to a situation in which current replacement rates are close to the optimal level.

ing insurance in the healthy state that yields payoff rY when injured is $prY/(1-p)$. If there is some insurance loading factor $h(h \geq 1)$ to cover administrative costs and a return to the insurance industry, then the cost becomes $hprY/(1-p)$.

The task of ascertaining the optimal insurance policy in the logarithmic utility function case is to

$$\underset{r}{\operatorname{Max}} V=(1-p)\alpha\ell n\{Y[1-(hpr/(1-p))](1-t)\}+p\ell n(Yr).$$

After substituting for the appropriate numerical values and taking the partial derivative with respect to r, one obtains the result that

$$\mathrm{hr}=0.85.$$

If workers' compensation were provided on an actuarially fair basis, the optimal replacement rate would be 85 percent. Less than full earnings replacement is desirable since the marginal utility of income is lower in the ill health state. Taking into account the role of taxes, an earnings replacement of 0.85 of gross earnings does replace most of the worker's after tax income.

Under the current workers' compensation system, administrative costs are nontrivial, so that after the insurance loading costs are taken into account workers receive 80¢ for each dollar contributed, or for each dollar of benefits they pay $1.25 in premiums (i.e., $h=1.25$). After taking these costs into account, the optimal replacement rate is 0.68.

The current workers' compensation formulas that provide for two-thirds wage replacement are close to optimal, given the role of administrative costs. The role of benefit caps and other provisions, however, reduces the effective replacement rate to only 0.64, which is slightly below this amount. In addition, if our reference point for benefit provision is what would be optimal if there were actuarially fair insurance available, then there is a much more substantial divergence from the optimal amount.

V. Conclusion

Analyses of risky decisions using market-based data are by necessity restricted to utilizing the information generated by the observed local tradeoff revealed in the market. Although this literature has yielded many profitable insights, the domain of inquiry has been substantially limited.

In this paper we explored the implications of knowing two wage-risk combinations along the individual's indifference map. This information was developed based on a survey of worker responses to the risks indicated by hazard warnings. The overall objective was to assess individuals' utility functions for good health and ill health, which will convey much more information about the character of individual preferences than the local tradeoff.

The two approaches that were used—a Taylor's series expansion with respect to a general functional form and a logarithmic utility function—each yielded similar results. Since being injured will clearly reduce the level of utility, the main question of interest is how the marginal utility of income is affected by an injury. In each case, the marginal utility of a given level of income was greater when healthy than when injured.

This result has fundamental implications for the optimal level of insurance since it implies that less than full insurance of income losses is optimal. This type of result has played a major role in the health economics and social insurance literature, but except in the case of death, the empirical foundation for making this determination has been lacking.

Even more striking is that the estimates of the logarithmic utility function enable us to ascertain not only whether less than full insurance is optimal but also what the optimal level of insurance is. In particular, we showed that the optimal earnings replacement rate for workers' compensation is 85 percent if insurance is provided on an actuarially fair basis and 68 percent if insurance is provided at the current degree of insurance loading. In each case, current benefit levels are slightly suboptimal, as has been shown using a different methodology by W.

Kip Viscusi and Michael J. Moore (1987), but in this case we can ascertain the extent of the divergence from optimality as well.

Knowledge of the utility function shape enables us to address a variety of other issues that have long been the subject of theoretical inquiry and empirical speculation. Perhaps the most striking result is the income elasticity of the value of an injury, which was found to range from 0.67 (Taylor's series) to roughly 1.0 (logarithmic case). This result enables one to make precise distinctions with respect to the heterogeneity in the value of risk-dollar tradeoffs across income groups. The greatest policy importance of this result is with respect to deferred risks since it indicates that the injury value figures used for deferred risk reductions should take into account the income growth of those affected by the regulations, leading to an adjustment that will serve to mute much of the role of discounting. These results may prove to be particularly useful in assessing the value of risk reduction to future generations.

There were two other types of concerns for which we obtained new results because of our focus on risk-dollar tradeoffs in more than a local region. As predicted by several theoretical analyses, increases in the base risk reduced the implicit value of an injury. The empirical sensitivity of the results to the base risk level was not, however, great.

Of much greater consequence was the change in the implicit value of an injury with the magnitude of any non-incremental risk change. Implicit values of an injury associated with the purchase of risk reductions diminished at an increasing rate as the extent of the risk reduction increased, and the implicit values associated with compensation for a risk increase rose at an increasing rate as the extent of the non-incremental risk change increased. As with the earlier results, these patterns are consistent with a rational economic choice model and lie outside the scope of concerns that can be addressed using market data.

Analysis of the survey experiment on worker responses to changes in their job risk has greatly expanded the range of risk-dollar tradeoffs that can be addressed empirically. The most reassuring aspect of the findings is that even the more refined predictions of the expected utility model with health state-dependent utility functions are borne out. Knowledge of the utility function shape also enables us to address for the first time many issues that have played a central role with respect of the economic performance and optimal government policies in contexts involving risk.

APPENDIX
SPECIFICATION TESTS

In this appendix we will summarize the results of the White-Domowitz (1984) specification tests. These tests have two objectives. First, they provide a formal test of the homoscedasticity assumption. Second, they provide a test of whether the model specification is correct up to an additive error term.

Consider first the logarithmic case. To perform the test, we must first define some terms. Let $\hat{\varepsilon}_i = \delta - f(x_i, \hat{\beta})$, and let $g_{ij}(\hat{\beta})$ be the jth element of the gradient $\partial f(X_i, \hat{\beta})/\partial\beta$, where the gradient is evaluated at the estimated parameter vector, $\hat{\beta}$. Define the vector ϕ_i to be formed by all nonredundant cross products of the gradient, $g_{ij}(\hat{\beta})g_{ik}(\hat{\beta})$, for $i, j \in (1,2,\dots, p)$. By definition, ϕ_i has a maximum length of $p(p+1)/2$. Let n equal the number of observations. The test statistic is generated from the regression of the square of the predicted residual on the vector ϕ_i and a constant,

$$\hat{\varepsilon}_i^2 = \gamma_0 + \phi_i\gamma,$$

where γ is a $p(p+1)/2\times 1$ vector of parameters to be estimated. The test statistic is formed by multiplying the number of observations times the (constant adjusted) R^2 of the above regression. The statistic is distributed as a χ^2 with $p(p+2)/2$ degrees of freedom. If nR^2 is less than the critical value of the χ^2 distribution, one accepts the null hypothesis of no heteroscedasticity.

The nR^2 test is also a test of the model specification. Rejection of the null hypothesis can be due either to heteroscedasticity or model misspecification. If one accepts the null hypothesis, then the model is correct up to an independent additive error term. In the first-order Taylor's series case, the nR^2 statistic of 1.469 for the full sample is well below the critical value of 7.81 at the 95 percent confidence level. The TNT and asbestos subsamples also have nR^2 values below the critical level, but the chloroacetophenone sample does not. Except for the chloroacetophenone subsample results, heteroscedasticity is not a problem, and there is also no evidence of statistically significant misspecification of the model.

The results for the logarithmic case, which are summarized at the bottom of Table 4, are similar. Consider the full sample results. The nR^2 statistics are 12.00 for Model 1 and 12.82 for Model 2, which is above the critical 95 percent confidence levels of 3.84 for Model 1 but below the critical level of 23.69 for Model 2. Similarly, the Model 1 results for asbestos are below the critical statistic for TNT and asbestos is not. Thus, one cannot reject either (i) the assumption of homoscedastic errors or (ii) the assumption that the model specification is correct up to an additive error term for Model 2 or for two subsample estimates of Model 1 (TNT and chloroacetophenone).

Although we cannot reject the hypothesis that both models are correctly specified, this result is not contradictory since we can consider the Taylor's series as simply approximating the logarithmic function. Additional evidence for this conclusion is in the closeness of the two estimates β_3 and $1/\alpha$.

REFERENCES

Arrow, Kenneth J., "Optimal Insurance and Generalized Deductibles," *Scandinavian Actuarial Journal*, 1974, 1–42. Reprinted in *Collected Papers of Kenneth J. Arrow, Volume 3: Individual Choice Under Certainty and Uncertainty*, Cambridge, MA: Harvard University Press, 1984, 212–60.

______, "The Role of Securities in the Optimal Allocation of Risk Bearing," *Review of Economic Studies*, April 1964, *3*, 91–96.

Cook, Philip J. and Graham, Daniel A., "The Demand for Insurance and Protection: The Case of Irreplaceable Commodities," *Quarterly Journal of Economics*, February 1977, *91*, 143–56.

Eisner, Robert and Strotz, Robert H., "Flight Insurance and the Theory of Choice," *Journal of Political Economy*, August 1961, *69*, 355–68.

Fuchs, Victor R. and Zeckhauser, Richard J., "Valuing Health—A Priceless Commodity," *American Economic Review Papers and Proceedings*, May 1987, *77*, 263–68.

Fuss, Melvyn A., McFadden, Daniel L. and Mundlak, Yair, "A Survey of Functional Form in the Economic Analysis of Production," in M. Fuss and D. McFadden, eds., *Production Economics: A Dual Approach to Theory and Applications*, Amsterdam: North-Holland, 1978, Vol. 1.

Gallant, A. Robert, "Nonlinear Regression," *American Statistician*, May 1975, *29*, 73–81.

Graham, Daniel A., "Cost-Benefit Analysis Under Uncertainty," *American Economic Review*, September 1981, *71*, 715–25.

Hirschleifer, Jack Z., *Investment, Interest, and Capital*, Englewood Cliffs, NJ: Prentice-Hall, 1970.

Karni, Edi, *Decision Making Under Uncertainty: The Case of State-Dependent Preferences*, Cambridge, MA: Harvard University Press, 1985.

Keeney, Ralph L. and Raiffa, Howard, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, New York: Wiley & Sons, 1976.

Kraus, Albert L. and Litzenberger, Robert H., "Market Equilibrium in a Multiperiod State Preference Model with Logarithmic Utility," *Journal of Finance*, December 1975, *30*, 1213–27.

Moore, Michael J. and Viscusi, W. Kip, *Compensation Mechanisms for Job Risks: Wages, Workers' Compensation, and Product Liability*, Princeton, NJ: Princeton University Press, forthcoming.

Newhouse, Joseph P. and Phelps, Charles E., "New Estimates of Price and Income Elasticities of Medical Care Services," in R. Rosett, ed., *The Role of Health Insurance in the Health Services Sector*, New York: National Bureau of Economic Research, 1976, 261–313.

Phelps, Charles E., *The Demand for Health Insurance: A Theoretical and Empirical Investigation*, Santa Monica, CA: Rand Corporation, 1973, Report R-1054-OEO.

______, "The Economics and Politics of Large-Scale Tax Reform: The Example of Employer-Paid Health Insurance Premiums," Working Paper, University of Rochester, 1987.

Pliskin, Joseph, Shepard, Donald S. and Weinstein, Milton C., "Utility Functions for Life Years and Health Status," *Operations Research*, 1980, *28*, 206–24.

Rubinstein, Mark E., "The Strong Case for the Generalized Logarithmic Utility Model as the Premier Model of Financial Markets," in H. Levy and M. Sarnat, eds., *Financial Decision Making Under Uncertainty*, New York: Academic Press, 1977.

Samuelson, Paul A., "Lifetime Portfolios Selection by Dynamic Stochastic Programming," *Review of Economics and Statistics*, August 1969, *51*, 239–46.

Shavell, Steven M., *Economic Analysis of Accident Law*, Cambridge, MA: Harvard

University Press, 1987.

Spence, A. Michael, "Consumer Misperceptions, Product Failure, and Producer Liability," *Review of Economic Studies,* October 1977, *44*, 561–72.

Torrance, George W., "Measurement of Health Status Utilities for Economic Appraisal: A Review," *Journal of Health Economics*, 1986, *5*, 1–30.

Viscusi, W. Kip, *Employment Hazards: An Investigation of Market Performance* Cambridge: Harvard University Press, 1979.

______, "Wealth Effects and Earnings Premiums for Job Hazards," *Review of Economics and Statistics*, August 1978, *60*, 3, 408–13.

______, "The Valuation of Risks to Life and Health: Guidelines for Policy Analysis," in J. Bentkover et al., eds., *Benefits Assessment: The State of the Art*, Dordrect: Reidel, 1986, 193–210.

______**Magat, Wesley A. and Huber, Joel C.,** "An Investigation of the Rationality of Consumer Valuations of Multiple Health Risks," *Rand Journal of Economics*, Winter 1987, *18*, 4, 465–79.

______ **and Moore, Michael J.,** "Workers' Compensation: Wage Effects, Benefit Inadequacies, and the Value of Health Losses," *Review of Economics and Statistics*, May 1987, *69*, 249–61.

______ **and O'Connor, Charles J.,** "Adaptive Responses to Chemical Labeling: Are Workers Bayesian Decision Makers?" *American Economic Review*, December 1984, *74*, 5, 942–56.

Weinstein, Milton C., Shepard, Donald S. and Pliskin, Joseph, "The Economic Value of Changing Morality Probabilities: A Decision-Theoretic Approach," *Quarterly Journal of Economics*, March 1980, *94*, 373–96.

White, Halbert L. and Domowitz, Ian, "Nonlinear Regressions with Dependent Observations," *Econometrica*, January 1984, *52*, 143–61.

Zeckhauser, Richard J., "Coverage for Catastrophic Illness," *Public Policy*, 1973, *21*, 149–72.

______, "Medical Insurance: A Case Study of the Tradeoff Between Risk Spreading and Appropriate Incentives," *Journal of Economic Theory*, March 1970, *2*, 1, 10–26.

Comparing Information in Forecasts from Econometric Models

By RAY C. FAIR AND ROBERT J. SHILLER*

The information contained in one model's forecast compared to that in another can be assessed from a regression of actual values on predicted values from the two models. We do this for forecasts of real GNP growth rates for different pairs of models. The models include a structural model (the Fair (1976) model), various versions of the vector autoregressive (VAR) model, and various versions of a model we call the "autoregressive components" (AC) model. Our procedure requires that forecasts make no use of future information, and we have been careful to try to insure this, including using the version of the Fair model that existed in 1976, the beginning of our test period. (JEL 132)

Many econometric models are used to forecast economic activity. These models differ in structure and in the data used, and so their forecasts are not perfectly correlated with each other. How should we interpret the differences in forecasts? Does each model have a strength of its own, so that each forecast represents useful information unique to it, or does one model dominate in the sense of incorporating all the information in the other models plus some?

Structural econometric models often make use of large information sets in forecasting a given variable. The information set used in a large-scale macroeconometric model is typically so large that the number of predetermined variables exceeds the number of observations available for estimating the model. Estimation can proceed effectively only because of the large number of a priori restrictions imposed on the model, restrictions that do not work out to be simple exclusion restrictions on the reduced form equation for the variable forecasted.

Vector autoregressive (VAR) models are typically much smaller than structural models and in this sense use less information. The above question with respect to VAR models versus structural models is thus whether the information not contained in VAR models (but contained in structural models) is useful for forecasting purposes. In other words, are the a priori restrictions of large-scale models useful in producing derived reduced forms that depend on so much information, or is most of the information extraneous?

One cannot answer this question by doing conventional tests of the restrictions in a structural model. These restrictions might be wrong in important ways and yet the model contain useful information. Even ignoring this point, however, one cannot perform such tests with most large-scale models because, as noted above, there are not enough observations to estimate unrestricted reduced forms.

We will examine the question whether one model's forecast of real GNP carries different information from another's by regressing the actual change in real GNP on the forecasted changes from the two models. This procedure, which is discussed in the next section, is related to the literature on encompassing tests[1] and the literature on the opti-

*Cowles Foundation, Box 2125, Yale Station, New Haven, CT 06520. We are indebted to the referees for many helpful comments, in particular for pointing out some of the special cases discussed at the end of Section I. This research was supported by the National Science Foundation under grant nos. SES-8520075 and SES-8717080.

[1] See, for example, Russell Davidson and James G. MacKinnon (1981), David F. Hendry and Jean-Francois Richard (1982), Yock Y. Chong and David F. Hendry

mal combination of forecasts.[2] From our point of view, this procedure has two advantages over the standard procedure of computing root mean squared errors (RMSEs) to compare alternative forecasts. First, if the RMSEs are close for two forecasts, little can be concluded about the relative merits of the two. With our procedure, as will be seen, one can sometimes discriminate more. Second, even if one RMSE is much smaller than the other, it may still be that the forecast with the higher RMSE contains information not in the other forecast. There is no way to test for this using the RMSE framework.

It should be stressed that our procedure does not allow us to discover whether all the variables in a model contribute useful information for forecasting. If, say, our regression results reveal that a large model contains all the information in smaller models plus some, it may be that the good results for the large model are due to a small subset of it. We can only say that the large model contains all the information in the smaller models that it has been tested against, not that it contains no extraneous variables.

We compare the forecasts from the structural model in Ray C. Fair (1976) with those of various atheoretical models. The atheoretical models include various versions of the VAR model and various versions of a model we call the "autoregressive components" (AC) model. The AC model divides GNP into components and estimates an autoregressive equation for each component. Like the Fair model, the AC model uses a fairly large amount of information on the components of GNP and yet it has a simple structure. Like parts of the Fair model, it uses lagged values of the components in the determination of the current values. The various versions of the AC model differ in their degree of disaggregation of the components, which allows us to examine how disaggregation affects the forecasts.

The AC model is of use to compare to the VAR model as well as to a structural model like the Fair model. It will be of interest to see within the class of atheoretical models how it compares to the VAR model. If, for example, it is found to contain information not in the VAR model, this indicates that the VAR model is missing useful information in the components. The AC model can be looked upon as an alternative to the VAR model when simple atheoretical models are desired.

Our procedure requires that forecasts be based only on information available prior to the forecast period. We call forecasts that meet this requirement "*quasi ex ante*" forecasts. To guard against future information creeping into the specification of the model and thus the forecasts, we chose the version of the Fair model that existed in 1976 and based all the tests on the period since 1976. We also converted the model to a model with no exogenous variables by adding an autoregressive equation for each exogenous variable in the model. Finally, we estimated the model (including the exogenous-variable equations) through period $t-1$ for each new beginning period t of the forecast. Using these "rolling" estimation forecasts is important because in doing so we are producing the actual forecasts that one could make with the model as time progresses.

We followed the same rolling estimation procedure for the VAR and AC models. These models have no exogenous variables, and so no adjustment was needed here. The Robert B. Litterman prior (1979) that we use for one of the versions of the VAR model was published near the beginning of the test period. The AC model is new, but it has such a simple structure that it is unlikely to be accused of having had future information used for its specification.

The quasi *ex ante* forecasts that we generate may have different properties from forecasts made with a model estimated with future data. If the model is misspecified (for example, parameters change through time), then the rolling estimation forecasts (where estimated parameters vary through time) may

(1986), and Grayham Mizon and Jean-Francois Richard (1986). See also Charles R. Nelson (1972) and J. Phillip Cooper and Charles R. Nelson (1975) for an early use of encompassing-like tests.

[2] See, for example, Clive W. J. Granger and Paul Newbold (1986).

carry rather different information from forecasts estimated over the entire sample.[3] Also, some models may use up more degrees of freedom in estimation than others, and with varied estimation procedures it is often very difficult to take formal account of the number of degrees of freedom used up. In the extreme case where there were so many parameters in a model that the degrees of freedom were completely used up when it was estimated (an obviously over parameterized model), it would be the case that the forecast value equals the actual value and there would be a spurious perfect correspondence between the variable forecasted and the forecast. One can guard against this degrees of freedom problem by requiring that no forecasts be within-sample forecasts.[4]

I. The Procedure

Let ${}_{t-s}\hat{Y}_{1t}$ denote a forecast of Y_t (in our application, log real gross national product at time t) made from model 1 using information available at time $t-s$ and using the model's estimation procedure and forecasting method each period. Let ${}_{t-s}\hat{Y}_{2t}$ denote the same thing for model 2. (In the notation above, these two forecasts are "*quasi ex ante*" forecasts.) The parameter s is the length ahead of the forecast, $s>0$. Note that the estimation procedure used to estimate a model and the model's forecasting method are considered by us as part of the model; we take no account of these procedures here. We consider the following regression equation:

$$(1)\quad Y_t - Y_{t-s} = \alpha + \beta\left({}_{t-s}\hat{Y}_{1t} - Y_{t-s}\right) + \gamma\left({}_{t-s}\hat{Y}_{2t} - Y_{t-s}\right) + u_t.$$

If neither model 1 nor model 2 contains any information useful for s-period-ahead forecasting of Y_t, then the estimates of β and γ should both be zero. In this case the estimate of the constant term α would be the average s-period-change in Y. If both models contain independent information[5] for s-period-ahead forecasting, then β and γ should both be nonzero. If both models contain information, but the information in, say, model 2 is completely contained in model 1 and model 1 contains further relevant information as well, then β but not γ should be nonzero. (If both models contain the same information, then the forecasts are perfectly correlated, and β and γ are not separately identified.)

The procedure we are proposing in this paper is to estimate equation (1) for different models' forecasts and test the hypothesis H_1 that $\beta=0$ and the hypothesis H_2 that $\gamma=0$. H_1 is the hypothesis that model 1's forecasts contain no information relevant to forecasting s periods ahead not in the constant term and in model 2, and H_2 is the hypothesis that model 2's forecasts contain no information not in the constant term and in model 1.

As we noted above, our procedure bears some relation to encompassing tests, but our setup and interests are somewhat different. For example, it does not make sense in our case to constrain β and γ to sum to one, as is usually the case for encompassing tests. If in our case both models' forecasts are just noise, the estimates of both β and γ should be zero. Also, say that the true process generating Y_t is Y_t equal to $X_t + Z_t$, where X_t and Z_t are independently distributed. Say that model 1 specifies that Y_t is a function of

[3] Even if the model is not misspecified, estimated parameters will change through time due to sampling error. If our purpose were to evaluate the forecasting ability of the *true* model (i.e., the model with the true coefficients), we would face a generated regressor problem. However, we are interested here in the performance of the model *and* its associated estimation procedure. If one were interested in adjusting for generated regressors, the correction discussed in Kevin M. Murphy and Robert H. Topel (1985) could not be directly applied here because the covariance matrix of the coefficient estimates used to generate the forecasts changes through time because of the use of the rolling regressions. Murphy and Topel require a single covariance matrix.

[4] Nelson (1972) and Cooper and Nelson (1975) do not require the forecasts to be based only on information through the previous period.

[5] If both models contain "independent information" in our terminology, their forecasts will not be perfectly correlated. This can arise either because the models use different data or because they use the same data but impose different restrictions on the reduced form.

X_t only and that model 2 specifies that Y_t is a function of Z_t only. Both forecasts should thus have coefficients of one in equation (1), and so in this case β and γ would sum to two. It also does not make sense in our setup to constrain the constant term α to be zero. If, for example, both models' forecasts were noise and we estimated equation (1) without a constant term, then the estimates of β and γ would not generally be zero when the mean of the dependent variable is nonzero.

It is also not sensible in our case to assume that u_t is identically distributed. It seems quite likely that u_t is heteroskedastic. If, for example, $\alpha=0$, $\beta=1$, and $\gamma=0$, u_t is simply the forecast error from model 1, and in general forecast errors are heteroskedastic. Also, we will be considering four-period-ahead forecasts in addition to one-period-ahead forecasts, and this introduces a third-order moving average process to the error term in equation (1).[6] We correct for both heteroskedasticity and the moving average process in the estimation of the standard errors of the coefficient estimates. We use the procedure given by Lars Peter Hansen (1982), Robert E. Cumby, John Huizinga, and Maurice Obstfeld (1983), and Halbert White and Ian Domowitz (1984) for the estimate of the asymptotic covariance matrix of the estimate $\hat{\theta}$ of the parameter vector $\theta \equiv (\alpha\ \beta\ \gamma)'$ in (1). Define X as the $T \times 3$ matrix of variables, whose row t is $X_t = [1, ({}_{t-s}\hat{Y}_{1t} - Y_{t-s}), ({}_{t-s}\hat{Y}_{2t} - Y_{t-s})]$, and let $\hat{u}_t = Y_t - Y_{t-s} - X_t\delta$. We estimate the covariance matrix of $\hat{\theta}, V(\hat{\theta})$, as:

$$(2) \qquad V(\hat{\theta}) = (X'X)^{-1}S(X'X)^{-1},$$

where

$$(3) \qquad S = \Omega_0 + \sum_{j=1}^{s-1} (\Omega_j + \Omega_j'),$$

$$(4) \qquad \Omega_j = \sum_{t=j+1}^{T} (u_t u_{t-j}) \hat{X}_t' \hat{X}_{t-j},$$

where $\hat{\theta}$ is the ordinary least squares estimate of θ and s is the forecast horizon. When s equals 1 the second term on the right hand side of (3) is zero, and the covariance matrix is simply Halbert White's (1980) correction for heteroskedasticity.

Note that as an alternative to equation (1) we could have regressed the *level* (rather than the change) of GNP on the forecasted *levels* and a constant. GNP has a very strong low frequency component. It may be an integrated process, and any sensible forecast of GNP will be cointegrated with GNP itself. The sum of β and γ will thus be constrained in effect to one, and in the levels regression we would be estimating in effect one less parameter. If GNP is an integrated process, running the levels regression with an additional independent variable Y_{t-1} (thereby estimating β and γ without constraining their sum to one) is essentially equivalent to our differenced regression (1).

It should finally be noted that there are cases in which an optimal forecast does not tend to be singled out as best in regressions of the form (1), even with many observations. Say the truth is $Y_t - Y_{t-1} = aX_{t-1} + e_t$. Say that model 1 does rolling regressions of $Y_t - Y_{t-1}$ on X_{t-1} and uses these regressions to forecast. Say that model 2 always takes the forecast to be bX_{t-1}, where b is some number other than a, so that model 2 remains forever an incorrect model. In equation (1) regressions the two forecasts tend to be increasingly collinear as time goes on; essentially they are collinear after the first part of the sample. Thus, the estimates of β and γ tend to be erratic. Adding a large number of observations does not cause the regressions to single out the first model; it only has the effect of enforcing that $\hat{\beta} + (\hat{\gamma}b)/a = 1$.

II. Quasi *Ex Ante* Forecasts

As noted above, we want forecasts from models that are based only on information through the period prior to the beginning of the forecast period (through period $t-s$ for a forecast for period t). There are four ways in which future information can creep into a current forecast. The first is if actual values

[6] The error term in equation (1) could, of course, be serially correlated even for the one-period-ahead forecasts. Such serial correlation does not appear to be a problem with any of the models we study here, however, and we have assumed it to be zero.

of the exogenous variables for periods after $t-s$ are used in the forecast. The second is if the coefficients of the model have been estimated over a sample period that includes observations beyond $t-s$. The third is if information beyond $t-s$ has been used in the specification of the model even though for purposes of the tests the model is only estimated through period $t-s$. The fourth is if information beyond period $t-s$ has been used in the revisions of the data for periods $t-s$ and back, such as revised seasonal factors and revised benchmark figures.

The VAR, AC, and AR models discussed below have no exogenous variables, and so there are no exogenous-variable problems for these models. The way we have handled the problem for the Fair model is to add autoregressive equations for the exogenous variables to the model. For each exogenous variable in the model an eighth-order autoregressive equation (with a constant term and time trend included) has been postulated. When these equations are added to the model, the model effectively has no exogenous variables in it. This method of dealing with exogenous variables in structural models was advocated by Cooper and Nelson (1975) and Stephen K. McNees (1981). McNees, however, noted that the method handicaps the model: "It is easy to think of exogenous variables (policy variables) whose future values can be anticipated or controlled with complete certainty even if the historical values can be represented by covariance stationary processes; to do so introduces superfluous errors into the model solution" (McNees, 1981, p. 404).

For the coefficient-estimate problem, we use rolling estimations for all the models. For the forecast for period t, we estimate the model through period $t-s$; for the forecast for period $t+1$, we estimate the model through period $t-s+1$; and so on. By "model" in this case we mean the model inclusive of any exogenous-variable equations. The beginning observation is held fixed for all the regressions; the sample expands by one observation each time a time period elapses.

The third problem—the possibility of using information beyond period $t-s$ in the specification of the model—is more difficult to handle. Models are typically changed through time, and model builders seldom go back to or are interested in "old" versions. We have, however, attempted to account for this problem in this paper regarding the Fair model. We consider the version of the Fair model that existed as of the second quarter of 1976.

We have done nothing about the data-revision problem in this paper. The data that have been used are the latest revised data. It would be extremely difficult to try to purge these data of the possible use of future information, and we have not tried. Note that it is not enough simply to use data that existed at any point in time (say period $t-s$) because data on the s-period-ahead value (period t) are needed to estimate equation (1). We would have to try to construct data for period t that are consistent with the old data for period $t-s$.

III. The Models

A. *The Fair Model* (*FAIR*)

The first version of the Fair model was presented in Fair (1976) along with the estimation method and method of forecasting with the model. This version was based on data through 1975 I. One important addition that was made to the model from this version was the inclusion of an interest rate reaction function in the model. This work is described in Ray C. Fair (1978), which is based on data through 1976 II. The version of the model in Fair (1976) consists of 26 structural stochastic equations, and with the addition of the interest rate reaction function, there are 27 stochastic equations. There are 106 exogenous variables, and for each of these variables an eighth-order autoregressive equation with a constant and time trend was added to the model. This gave a model of 133 stochastic equations, and this is the version that was used.

For the rolling estimations, the first estimation period ended in 1976 II, which is the first quarter in which the model could definitely be said to exist. This allowed the model to be estimated 40 times (through 1986 I).

Because the version of the Fair model used here existed as of 1976 II, because it has in effect no exogenous variables, and because it is estimated via rolling estimation, the forecasts from it can be said to be forecasts that are truly based only on information through the period prior to the first period of the forecast (except for the data revision problem).[7] This may be the first time that a large model this old has been tested.

B. *The VAR Models (VAR4, VAR4P1, VAR4P2, VAR4P3, VAR1, and VAR2)*

We consider six VAR models in this paper. The first, VAR4, estimated by ordinary least squares, is the same as the model used in Christopher A. Sims (1980) except that we have added the three-month Treasury bill rate to the model. There are seven variables in the model: real GNP, the GNP deflator, the unemployment rate, the nominal wage rate, the price of imports, the money supply, and the bill rate. All but the unemployment rate and the bill rate are in logs. Each equation consists of each variable lagged one through four times, a constant, and a time trend, for a total of 30 coefficients to estimate.

The next three VAR models—VAR4P1, VAR4P2, and VAR4P3—have Bayesian priors imposed on the coefficients of VAR4. We impose the Litterman prior that the variables follow univariate random walks. The standard deviations of the prior take the form

$$(5)\quad S(i,j,k)=\gamma g(k)f(i,j)(s_j/s_i),$$

where i indexes the left-hand side variable, j indexes the right-hand side variables, and k indexes the lag. s_i is the standard error of the unrestricted equation for variable i. VAR4P1 imposes parameter values that imply fairly loose priors. They are: 1) $f(i,j)=1$ for all i and j, 2) $g(k)=1$ for all k, and 3) $\gamma=0.2$. VAR4P2 imposes parameter values that imply much tighter priors: 1) $f(i,j)=1$ for $i=j$, $f(i,j)=0.5$ for $i\neq j$, 2) $g(k)=k^{-1}$, and 3) $\gamma=0.1$. VAR4P3 is the same as VAR4P2 except that $f(i,j)=0.2$ for $i\neq j$, which implies even tighter priors than for VAR4P2. The parameter values for VAR4P2 are those imposed by Litterman (1979, p. 49).

The fifth VAR model, VAR2, uses only the first two lags of each variable, for a total of 16 coefficients in each equation. The sixth model, VAR1, uses only each variable lagged once, for a total of 9 coefficients. No priors were imposed on VAR2 and VAR1; they were estimated by ordinary least squares.

Each VAR model was estimated 40 times using the same sample periods as were used for the Fair model. Each model was then used to make 40 forecasts of real GNP.

C. *The AC Models (AC-6, AC-13, AC-17, AC-48, AC-E6, AC-E13, AC-E17, and AC-E48)*

Time-series models like VAR models typically ignore the components of GNP. For example, the VAR models used in this paper contain no components. The current model used by Christopher A. Sims (serial) for forecasting includes only the component nonresidential fixed investment. Including many components in a VAR model rapidly uses up degrees of freedom, and this is undoubtedly one of the main reasons the components are seldom used. A possible alternative to the VAR approach, but one that also does not use much economic theory, is to

[7] This statement needs to be qualified slightly. Although the structural stochastic equations used for the Fair model are exactly as in Fair (1976) and (1978)—same left-hand side and right-hand side variables and same functional forms—the data revisions in the National Income Accounts since 1976 have required slight modifications to some of the identities in the model. Also, the identities in Fair (1976) for the government sector are for the total government sector, whereas in the version used here there are separate identities for the federal government sector and the state and local government sector. This disaggregation of the government sector does not affect anything except that it means that there are more exogenous variables (and thus more exogenous-variable equations) in the version used here than there were in Fair (1976). No structural stochastic equations need to be modified because of this disaggregation because the government variables do not appear as explanatory variables in these equations. The government variables appear only in the identities.

model each of the components of real GNP by a simple autoregressive equation (but not real GNP itself) and then determine real GNP as the sum of the components, (i.e., by the GNP identity).

The AC model may be regarded as an extreme caricature of large-scale macroeconometric models. Like the latter, lagged values of components are used to predict components, and the predicted components are added up to predict GNP. The AC model, however, treats all components in a simple and symmetrical way and carries the number of lags further than is often the case with large-scale models. Note also that the AC model could be adapted to forecast variables other than GNP. For any variable that is determined by an identity, autoregressive equations could be used to forecast the right-hand side variables and then the identity used to forecast the variable itself. Also, the AC model for real GNP could be added to equations explaining other variables to create a more complete macroeconometric model.

We have considered eight AC models in this paper, all estimated by ordinary least squares. The models are first distinguished by whether they include 6, 13, 17, or 48 components. Increasing the disaggregation of the components allows one to examine how much additional useful information for forecasting purposes is contained in the more disaggregated components. Each equation for a component contains the first eight lagged values of the component, a constant, and a time trend. None of the AC equations is in log form. For the AC-6, AC-13, AC-17, and AC-48 models, these are all the variables included in the equations. For the AC-E6, AC-E13, AC-E17, and AC-E48 models (E for "extended"), the first four lagged values of real GNP are added to each equation. This allows for an impact of aggregate economic activity on each component and uses up only four degrees of freedom per equation. The components used for each model are listed in the Appendix. The 17 component models (AC-17 and AC-E17) use the same components as does the Fair model. The same sample periods and procedures were used for the AC models as were used for the Fair and VAR models except the for the 6, 13, and 48 component models the beginning quarter for the estimation periods was 1961 I rather than 1954 I.[8]

The AC models are of interest in two respects. First, if the Fair model turns out to dominate the VAR models (which it does), it is of interest to know if this is due simply to the fact that the Fair model is dealing with the lagged components of GNP. If it is as simple as this, then the AC models might do even better, and this can be tested. Second, the AC models are to some extent competitors of the VAR models within the class of atheoretical models, at least regarding the predictions of GNP. Both models are based on very little economic theory. It is thus of interest to see if one type of model dominates the other.

D. *The Autoregressive Models (AR4 and AR8)*

AR4 and AR8 are simple benchmark models, estimated by ordinary least squares. For AR4 real GNP was regressed on its first four lagged values, a constant, and a time trend. For AR8 real GNP was regressed on its first eight lagged values, a constant, and a time trend. The same sample periods were used here as were used for the Fair and VAR models.

IV. The Results

The results comparing the Fair model to the other models are presented in Tables 1 and 2. The sample period used for the one-quarter-ahead results is 1976 III–1986 II, for a total of 40 observations. The sample period for the four-quarter-ahead results is 1972

[8] This choice was dictated by the available data. Fortunately, the results do not appear to be sensitive to the use of the later beginning quarter. For the 17 component model the estimation periods could begin in 1954 I, and so two versions of AC-17 and AC-E17 were estimated, one for the periods beginning in 1954 I and one for the periods beginning in 1961 I. The two versions gave very similar results.

TABLE 1—BIAS AND RMSE FOR EACH MODEL'S FORECAST

	One-Quarter-Ahead Forecasts Sample Period =1976 III–1986 II				Four-Quarter-Ahead Forecasts Sample Period =1977 II–1986 II			
	const	SE	DW	RMSE	const	SE	DW	RMSE
FAIR	0.0034 (2.43)	0.00883	2.04	0.00945	0.0087 (3.60)	0.0146	1.04	0.0170
VAR1	−0.0014 (0.78)	0.01179	1.47	0.01188	−0.0127 (2.93)	0.0263	0.56	0.0292
VAR2	−0.0021 (1.23)	0.01090	1.36	0.01111	−0.0157 (3.33)	0.0285	0.40	0.0326
VAR4	0.0012 (0.68)	0.01145	1.22	0.01152	−0.0011 (0.24)	0.0291	0.53	0.0291
VAR4P1	0.0004 (0.24)	0.01105	1.41	0.01106	−0.0028 (0.59)	0.0291	0.43	0.0292
VAR4P2	−0.0017 (1.07)	0.01004	1.68	0.01018	−0.0105 (2.90)	0.0220	0.52	0.0244
VAR4P3	−0.0010 (0.66)	0.00976	1.75	0.00981	−0.0063 (1.65)	0.0231	0.43	0.0239
AC-6	0.0016 (0.93)	0.01070	1.64	0.01081	0.0052 (1.26)	0.0251	0.42	0.0256
AC-13	0.0007 (0.39)	0.01071	1.49	0.01073	0.0007 (0.18)	0.0259	0.38	0.0259
AC-17	0.0018 (1.06)	0.01080	1.56	0.01095	0.0031 (0.72)	0.0264	0.38	0.0266
AC-48	0.0008 (0.46)	0.01141	1.18	0.01144	0.0019 (0.41)	0.0281	0.31	0.0282
AC-E6	0.0020 (1.23)	0.01047	1.96	0.01066	0.0068 (1.69)	0.0245	0.54	0.0254
AC-E13	0.0005 (0.34)	0.00957	1.93	0.00958	0.0001 (0.02)	0.0222	0.52	0.0222
AC-E17	0.0017 (1.18)	0.00936	2.25	0.00952	0.0035 (0.95)	0.0228	0.62	0.0231
AC-E48	0.0009 (0.60)	0.00993	1.84	0.00997	0.0037 (0.84)	0.0266	0.50	0.0268
AR4	0.0026 (1.60)	0.01023	2.02	0.01055	0.0134 (2.94)	0.0277	0.48	0.0308
AR8	0.0028 (1.63)	0.01067	1.98	0.01102	0.0147 (3.13)	0.0286	0.52	0.0321

Notes: const is the estimate of the constant term in a regression of the forecast error on the constant term. The forecast error is ${}_{t-1}\hat{Y}_{1t} - Y_t$ for the one-quarter-ahead results and ${}_{t-4}\hat{Y}_{1t} - Y_t$ for the four-quarter-ahead results. Y is the log of real GNP.
t-statistics in absolute value are in parentheses.
SE is the estimated standard error of the regression not adjusting for degrees of freedom.
RMSE is the root mean squared error of the forecast.

II–1986 II, for a total of 37 observations. Remember that each observation for a model's forecast is based on a different set of coefficient estimates of the model—the rolling estimation. Remember also that for the Fair model all exogenous variable values are generated from the autoregressive equations; no actual values are used. Finally, remember that in Table 2 the estimated standard errors of the coefficient estimates are corrected for heteroskedasticity and (for the four-quarter-ahead results) for the moving average process of the error term.[9]

[9] In two cases for the four-quarter-ahead results the matrix $V(\hat{\theta})$ was singular or nearly singular. In these two cases we assumed a second-order MA process for the error term instead of a third order, which solved the problem. Had this been a more widespread problem, we would have used one of the estimators in Donald W. K. Andrews (1987), but this seemed unnecessary given only two failures. The two failures are FAIR versus VAR4 and FAIR versus VAR4P1.

TABLE 2—FAIR MODEL VERSUS THE OTHERS: ESTIMATES OF EQUATION (1)

	One-Quarter-Ahead Forecasts					Four-Quarters-Ahead Forecasts				
	Dependent Variable is $Y_t - Y_{t-1}$					Dependent Variable is $Y_t - Y_{t-4}$				
	Sample Period = 1976 III–1986 II					Sample Period = 1977 II–1986 II				
Other		FAIR	OTHER				FAIR	OTHER		
Model	const	$_{t-1}\hat{Y}_{1t} - Y_{t-1}$	$_{t-1}\hat{Y}_{2t} - Y_{t-1}$	SE	DW	const	$_{t-4}\hat{Y}_{1t} - Y_{t-4}$	$_{t-4}\hat{Y}_{2t} - Y_{t-4}$	SE	DW
VAR1	−0.0034	1.05	−0.10	0.00915	2.02	−0.0135	1.06	0.20	0.0134	1.66
	(1.20)	(4.51)	(0.38)			(2.28)	(6.78)	(3.09)		
VAR2	−0.0029	0.89	0.14	0.00912	2.02	−0.0135	1.08	0.18	0.0135	1.59
	(1.00)	(3.47)	(0.69)			(2.08)	(6.30)	(2.66)		
VAR4	−0.0031	0.86	0.14	0.00908	1.93	−0.0164	1.07	0.20	0.0130	1.62
	(1.06)	(3.53)	(0.89)			(2.72)	(6.53)	(3.31)		
VAR4P1	−0.0033	0.91	0.11	0.00914	2.02	−0.0159	1.08	0.19	0.0132	1.64
	(1.07)	(3.83)	(0.70)			(2.55)	(6.32)	(2.79)		
VAR4P2	−0.0031	0.89	0.17	0.00913	2.06	−0.0145	1.04	0.29	0.0133	1.61
	(1.05)	(3.56)	(0.64)			(2.47)	(6.20)	(2.76)		
VAR4P3	−0.0041	0.87	0.37	0.00910	2.08	−0.0209	1.07	0.49	0.0135	1.57
	(1.15)	(3.73)	(0.89)			(3.69)	(6.70)	(2.24)		
AC-6	−0.0041	0.94	0.16	0.00913	2.10	−0.0201	1.13	0.22	0.0140	1.50
	(1.11)	(3.98)	(0.67)			(3.09)	(6.89)	(1.43)		
AC-13	−0.0034	0.99	0.02	0.00918	2.04	−0.0193	1.18	0.16	0.0144	1.45
	(0.92)	(3.93)	(0.05)			(2.39)	(7.34)	(0.58)		
AC-17	−0.0031	1.00	−0.04	0.00917	2.03	−0.0164	1.22	−0.00	0.0145	1.42
	(0.84)	(4.07)	(0.14)			(1.86)	(8.11)	(0.00)		
AC-48	−0.0039	0.97	0.10	0.00915	2.04	−0.0191	1.19	0.14	0.0143	1.48
	(1.15)	(4.00)	(0.50)			(2.85)	(7.63)	(0.85)		
C-E6	−0.0042	0.91	0.20	0.00910	2.14	−0.0195	1.15	0.17	0.0142	1.51
	(1.23)	(3.69)	(0.80)			(2.80)	(7.05)	(1.15)		
AC-E13	−0.0043	0.81	0.40	0.00899	2.16	−0.0171	1.17	0.10	0.0145	1.41
	(1.35)	(2.81)	(1.05)			(2.47)	(6.30)	(0.41)		
AC-E17	−0.0045	0.76	0.42	0.00894	2.31	−0.0168	1.20	0.04	0.0145	1.42
	(1.41)	(2.79)	(1.50)			(2.50)	(6.19)	(0.16)		
AC-E48	−0.0047	0.86	0.36	0.00894	2.21	−0.0168	1.21	0.03	0.0145	1.43
	(1.42)	(3.64)	(1.65)			(2.62)	(7.16)	(0.19)		
AR4	−0.0084	0.96	0.59	0.00885	2.47	−0.0168	1.22	0.01	0.0145	1.42
	(2.01)	(4.51)	(1.87)			(1.12)	(9.24)	(0.03)		
AR8	−0.0067	0.98	0.38	0.00901	2.32	−0.0144	1.22	−0.05	0.0145	1.41
	(1.72)	(4.48)	(1.33)			(1.70)	(9.54)	(0.27)		

Notes: Y = log of real GNP.
t-statistics in absolute value are in parentheses.
See text for discussion of estimation methods.

The bias and RMSE for each forecast are presented in Table 1.[10] The bias is estimated by regressing the forecast error (predicted change minus actual change) on a constant. If the constant term is zero in this regression then the standard error (SE) of the regression is the same as the RMSE; otherwise the RMSE is larger. The errors are roughly in percentage points (0.01 is a 1 percent error) because real GNP is in logs. For the one-quarter-ahead results the Fair model has the largest bias, but even this bias is only 0.34 percent. It also has the smallest RMSE, although a number of RMSEs are quite close. The second best model in terms of RMSE is AC-E17. The best VAR model is VAR4P3. For the four-quarter-ahead results a number of the estimated biases are significant. The Fair model has by far the smallest RMSE. The best VAR model is again VAR4P3, and the best AC model is now AC-E13. AC-E13 and AC-E17 are better than VAR4P3.

[10] For reference purposes, the predicted values and errors for three models—Fair, VAR4P3, and AC-E17—are presented in Table A in the appendix.

Consider now the results in Table 2. The coefficient estimate for the Fair model forecast is always significant at the 5 percent level for both the one-quarter-ahead and four-quarter-ahead results. None of the coefficient estimates for the other models' forecasts is significant at this level for the one-quarter-ahead results. The regression gives a fairly large weight to the AR4 and AC-E17 forecasts, although these are not quite statistically significant. For the four-quarter-ahead results the only significant estimates (aside from those for the Fair model) are for the VAR models. The results across the different VAR models are fairly close, with perhaps VAR4 performing the best.

The results in Table 2 are thus rather striking. They provide strong support for the hypothesis that the Fair model carries useful information not in the other models. The significance of the VAR forecasts for the four-quarter-ahead results indicates that some information is in the VAR forecasts that the Fair model is not using for the four-quarter-ahead forecasts, but this is the only significantly negative aspect of the results for the Fair model. It perhaps indicates some dynamic misspecification for the Fair model.

Comparing Tables 1 and 2 shows one of the advantages of our procedure over the RMSE procedure. For the one-quarter-ahead results in Table 1 the RMSEs are all fairly close, and even though the RMSE is smallest for the Fair model, it is only slightly smaller. One might conclude from this table that the models are all about the same. The results in Table 2, on the other hand, show that the Fair model dominates the others by a fairly large amount. Conversely, the four-quarter-ahead results in Table 1 show the Fair model dominating the others by a large amount, but the results in Table 2 show that the VAR forecasts contain information not in the Fair forecasts. One would not have known this from Table 1.

Table 3 goes on to compare the VAR models with the AC and AR models. It is hard from Table 2 to pick out which AC and VAR model performs the best because the models are so dominated by the Fair model, but Table 3 provides more ability to discriminate. Regarding the AC models, the best results in terms of significant coefficient estimates are obtained for the 13 and 17 component versions and for the extended (AC-E) versions. In other words, going from 6 to 13 or 17 components does help, but going beyond this does not seem to add further useful information, and adding the lagged values of real GNP to the equations seems to add useful information.

The best performing AC model in Table 3 is probably AC-E17, and so consider the comparisons of AC-E17 with the VAR models. AC-E17 performs better than any VAR model for the one-quarter-ahead results. No VAR model coefficient estimate is significant for these comparisons, although some are close to being significant. For the four-quarter-ahead results the VAR models perform about as well as does AC-E17. The best fit is for AC-E17 versus VAR4P2, where the t-statistic for VAR4P2 is 3.73 and the t-statistic for AC-E17 is 3.41. In other words, both AC-E17 and the VAR models appear to contain independent information useful for forecasting four quarters ahead. Comparing across the VAR models, VAR4P2 and VAR4P3 are probably the best, although the results are quite close across the models. The results in Table 3 also show that the VAR models dominate the AR models. Clearly the VAR models contain information not in the AR models, but not vice versa.

The results in Table 3 thus indicate that AC models like AC-E17 contain useful forecasting information not contained in even the best VAR model. In other words, there appears to be useful forecasting information in the components of GNP that is not captured in the VAR models, and so within the class of fairly atheoretical models, AC models appear to be useful alternatives to the VAR models. This conclusion is strengthened by the fact that for the one-quarter-ahead forecasts the VAR models appear to contain little information not already in AC-E17.

Regarding the AC models, note from Table 2 that the AC forecasts are not statistically significant at the 5 percent level when compared with the Fair forecasts. Although the AC-E17 forecast gets a weight of 0.42 for the

TABLE 3—VAR MODELS VERSUS AC AND AR MODELS: ESTIMATES OF EQUATION (1)
C = CONSTANT, V = VAR MODEL, A = AC OR AR MODEL
ONE-QUARTER-AHEAD FORECASTS   DEPENDENT VARIABLE IS $Y_t - Y_{t-1}$ SAMPLE PERIOD = 1976 III–1986 II

	VAR1			VAR2			VAR4			VAR4P1			VAR4P2			VAR4P3		
	C	*V*	*A*	*C*	*V*	*A*	*C*	*V*	*A*	*C*	*V*	*A*	*C*	*V*	*A*	*C*	*V*	*A*
AC-6	0.0029	0.14	0.37	0.0034	0.39	0.17	0.0019	0.36	0.23	0.0022	0.36	0.23	0.0025	0.56	0.16	−0.0004	0.94	0.21
	(0.98)	(0.59)	(1.94)	(1.20)	(2.18)	(0.74)	(0.62)	(2.05)	(0.97)	(0.70)	(2.21)	(0.99)	(0.83)	(2.24)	(0.71)	(0.10)	(2.39)	(0.93)
		[0.01057]			[0.01014]			[0.01003]			[0.01020]			[0.01011]			[0.01008]	
AC-13	0.0035	0.21	0.28	0.0039	0.43	0.10	0.0023	0.38	0.18	0.0023	0.40	0.21	0.0028	0.60	0.11	−0.0004	1.01	0.19
	(0.95)	(0.94)	(0.74)	(1.12)	(2.39)	(0.25)	(0.66)	(2.13)	(0.44)	(0.64)	(2.37)	(0.51)	(0.78)	(2.38)	(0.29)	(0.10)	(2.52)	(0.49)
		[0.01068]			[0.01018]			[0.01008]			[0.01024]			[0.01013]			[0.01011]	
AC-17	0.0026	0.25	0.32	0.0027	0.43	0.23	0.0014	0.39	0.25	0.0011	0.42	0.31	0.0016	0.61	0.24	−0.0018	1.05	0.30
	(0.67)	(1.05)	(1.05)	(0.74)	(2.49)	(0.75)	(0.39)	(2.31)	(0.82)	(0.29)	(2.60)	(1.01)	(0.41)	(2.50)	(0.80)	(0.38)	(2.66)	(0.99)
		[0.01065]			[0.01013]			[0.01005]			[0.01019]			[0.01008]			[0.01005]	
AC-48	0.0039	0.23	0.20	0.0040	0.43	0.10	0.0026	0.39	0.13	0.0027	0.40	0.14	0.0025	0.60	0.16	−0.0008	1.04	0.22
	(1.48)	(1.00)	(1.09)	(1.69)	(2.38)	(0.50)	(1.00)	(2.21)	(0.69)	(1.06)	(2.43)	(0.73)	(0.92)	(2.48)	(0.81)	(0.22)	(2.68)	(1.14)
		[0.01067]			[0.01017]			[0.01007]			[0.01024]			[0.01009]			[0.01004]	
AC-E6	0.0019	0.14	0.48	0.0022	0.36	0.32	0.0012	0.33	0.33	0.0014	0.32	0.34	0.0014	0.50	0.31	−0.0011	0.87	0.33
	(0.57)	(0.55)	(1.89)	(0.69)	(1.91)	(1.31)	(0.39)	(1.81)	(1.27)	(0.45)	(1.81)	(1.29)	(0.44)	(1.96)	(1.27)	(0.27)	(2.19)	(1.39)
		[0.01037]			[0.01000]			[0.00992]			[0.01009]			[0.00997]			[0.00995]	
AC-E13	0.0000	0.13	0.82	0.0003	0.31	0.69	−0.0007	0.30	0.70	−0.0005	0.28	0.71	−0.0003	0.42	0.67	−0.0023	0.73	0.68
	(0.03)	(0.58)	(2.54)	(0.09)	(1.88)	(2.12)	(0.21)	(1.87)	(1.99)	(0.14)	(1.86)	(2.06)	(0.07)	(1.85)	(2.03)	(0.57)	(2.05)	(2.10)
		[0.00988]			[0.00958]			[0.00946]			[0.00963]			[0.00958]			[0.00957]	
AC-E17	−0.0008	0.09	0.83	−0.0005	0.27	0.70	−0.0015	0.28	0.71	−0.0011	0.24	0.72	−0.0009	0.36	0.96	−0.0026	0.62	0.69
	(0.22)	(0.40)	(3.26)	(0.15)	(1.64)	(2.89)	(0.47)	(1.82)	(2.66)	(0.35)	(1.67)	(2.73)	(0.27)	(1.54)	(2.74)	(0.67)	(1.74)	(2.87)
		[0.00967]			[0.00944]			[0.00930]			[0.00949]			[0.00946]			[0.00945]	
AC-E48	0.0011	0.14	0.64	0.0013	0.32	0.51	0.0003	0.31	0.51	0.0006	0.29	0.52	0.0004	0.47	0.51	−0.0019	0.82	0.52
	(0.35)	(0.63)	(3.29)	(0.50)	(1.85)	(2.63)	(0.11)	(1.82)	(2.36)	(0.21)	(1.83)	(2.51)	(0.15)	(2.05)	(2.75)	(0.50)	(2.27)	(2.88)
		[0.01009]			[0.00978]			[0.00967]			[0.00985]			[0.00971]			[0.00969]	
AR4	−0.0003	0.21	0.63	0.0004	0.39	0.48	−0.0021	0.38	0.62	−0.0005	0.36	0.49	0.0004	0.54	0.39	−0.0014	0.92	0.31
	(0.07)	(0.88)	(1.65)	(0.11)	(2.17)	(1.31)	(0.52)	(2.44)	(1.69)	(0.11)	(2.17)	(1.26)	(0.10)	(2.03)	(1.00)	(0.32)	(2.03)	(0.76)
		[0.01044]			[0.01001]			[0.00978]			[0.01010]			[0.01003]			[0.01008]	
AR8	0.0021	0.23	0.36	0.0025	0.42	0.24	0.0003	0.39	0.35	0.0017	0.40	0.23	0.0023	0.59	0.15	0.0002	1.05	0.06
	(0.50)	(0.94)	(0.96)	(0.63)	(2.31)	(0.67)	(0.08)	(2.36)	(0.97)	(0.43)	(2.28)	(0.60)	(0.53)	(3.10)	(0.55)	(0.06)	(2.29)	(0.15)
		[0.01063]			[0.01013]			[0.00999]			[0.01023]			[0.01012]			[0.01014]	
AC-6	0.0124	0.45	0.25	0.0127	0.37	0.31	0.0073	0.36	0.32	0.0082	0.36	0.30	0.0095	0.69	0.19	−0.0052	1.03	0.34
	(1.27)	(1.88)	(0.59)	(1.28)	(1.61)	(0.74)	(0.82)	(1.91)	(0.79)	(0.91)	(1.71)	(0.73)	(1.05)	(1.87)	(0.43)	(0.49)	(1.57)	(0.86)
		[0.0230]			[0.0236]			[0.0236]			[0.0237]			[0.0222]			[0.0230]	
AC-13	0.0142	0.51	0.19	0.0164	0.44	0.19	0.0100	0.43	0.21	0.0107	0.43	0.21	0.0110	0.75	0.13	−0.0059	1.17	0.32
	(1.04)	(2.34)	(0.32)	(1.13)	(2.00)	(0.30)	(0.79)	(2.32)	(0.33)	(0.82)	(2.15)	(0.32)	(0.86)	(2.27)	(0.21)	(0.50)	(1.90)	(0.55)
		[0.0232]			[0.0240]			[0.0240]			[0.0240]			[0.0223]			[0.0234]	
AC-17	0.0049	0.50	0.49	0.0079	0.44	0.47	0.0020	0.42	0.46	0.0024	0.43	0.47	0.0016	0.72	0.45	−0.0155	1.22	0.59
	(0.42)	(3.66)	(1.18)	(0.64)	(2.93)	(1.04)	(0.17)	(3.63)	(1.03)	(0.20)	(3.38)	(1.04)	(0.14)	(3.35)	(1.06)	(1.22)	(2.87)	(1.39)
		[0.0225]			[0.0235]			[0.0235]			[0.0235]			[0.0218]			[0.0226]	
AC-48	0.0199	0.56	−0.04	0.0234	0.52	−0.09	0.0163	0.51	−0.10	0.0170	0.51	−0.09	0.0155	0.82	−0.07	−0.0013	1.28	0.07
	(1.77)	(2.48)	(0.09)	(1.93)	(2.18)	(0.18)	(1.55)	(2.56)	(0.20)	(1.59)	(2.31)	(0.18)	(1.49)	(2.40)	(0.16)	(0.11)	(1.97)	(0.15)
		[0.0233]			[0.0219]			[0.0241]			[0.0241]			[0.0224]			[0.0237]	
AC-E6	0.0087	0.42	0.36	0.0092	0.35	0.40	0.0044	0.34	0.40	0.0051	0.34	0.40	0.0064	0.64	0.30	−0.0075	0.99	0.42
	(0.78)	(2.53)	(1.19)	(0.79)	(2.14)	(1.40)	(0.39)	(2.49)	(1.37)	(0.45)	(2.27)	(1.33)	(0.59)	(2.38)	(0.93)	(0.61)	(1.96)	(1.41)
		[0.0226]			[0.0232]			[0.0233]			[0.0233]			[0.0219]			[0.0226]	
AC-E13	0.0003	0.39	0.78	0.0016	0.33	0.81	−0.0027	0.31	0.80	−0.0025	0.32	0.81	−0.0020	0.57	0.73	−0.0145	0.96	0.82
	(0.03)	(4.02)	(3.17)	(0.13)	(2.69)	(2.92)	(0.22)	(3.35)	(2.77)	(0.21)	(3.25)	(2.86)	(0.19)	(3.80)	(3.19)	(1.14)	(3.12)	(3.70)
		[0.0199]			[0.0207]			[0.0209]			[0.0207]			[0.0195]			[0.0199]	
AC-E17	−0.0015	0.42	0.74	−0.0003	0.36	0.76	−0.0049	0.34	0.76	−0.0044	0.34	0.76	−0.0036	0.60	0.69	−0.0171	1.01	0.77
	(0.14)	(3.97)	(3.52)	(0.03)	(2.81)	(3.15)	(0.42)	(3.43)	(2.95)	(0.39)	(3.31)	(3.00)	(0.37)	(3.73)	(3.41)	(1.47)	(3.20)	(3.81)
		[0.0201]			[0.0209]			[0.0210]			[0.0210]			[0.0197]			[0.0202]	
AC-E48	0.0111	0.47	0.29	0.0131	0.40	0.30	0.0081	0.39	0.28	0.0083	0.39	0.29	0.0078	0.69	0.25	−0.0064	1.12	0.34
	(1.24)	(3.35)	(1.20)	(1.40)	(2.66)	(1.19)	(0.87)	(3.05)	(1.06)	(0.91)	(2.90)	(1.08)	(0.87)	(2.99)	(0.98)	(0.55)	(2.48)	(1.34)
		[0.0226]			[0.0234]			[0.0236]			[0.0235]			[0.0219]			[0.0227]	
AR4	0.0075	0.54	0.29	0.0069	0.48	0.36	−0.0017	0.46	0.41	0.0018	0.47	0.34	0.0106	0.78	0.09	−0.0044	1.31	0.11
	(0.36)	(3.94)	(0.58)	(0.34)	(3.32)	(0.68)	(0.08)	(4.15)	(0.74)	(0.08)	(3.77)	(0.59)	(0.48)	(3.70)	(0.16)	(0.19)	(2.71)	(0.16)
		[0.0232]			[0.0240]			[0.0240]			[0.0240]			[0.0224]			[0.0237]	
AR8	0.0212	0.55	−0.06	0.0162	0.49	0.12	0.0075	0.47	0.17	0.0124	0.47	0.07	0.0218	0.80	−0.20	0.0086	1.36	−0.23
	(1.13)	(4.08)	(0.14)	(0.95)	(3.38)	(0.32)	(0.41)	(4.13)	(0.42)	(0.65)	(3.78)	(0.16)	(1.23)	(3.97)	(0.49)	(0.46)	(2.89)	(0.47)
		[0.0233]			[0.0241]			[0.0241]			[0.0241]			[0.0223]			[0.0236]	

Notes: Y = log of real GNP.
t-statistics in absolute value are in parentheses.
Estimated standard errors of the regressions are in brackets.
See text for discussion of estimation methods.
$V = {}_{t-1}\hat{Y}_{1t} - Y_{t-1}$ for one-quarter-ahead results for VAR models.
$V = {}_{t-y}\hat{Y}_{1t} - Y_{t-4}$ for four-quarter-ahead results for VAR models.
$A = {}_{t-1}\hat{Y}_{2t} - Y_{t-1}$ for one-quarter-ahead results for AC and AR models.
$A = {}_{t-4}\hat{Y}_{2t} - Y_{t-4}$ for four-quarter-ahead results for AC and AR models.

one-quarter-ahead results in Table 2, it is not statistically significant, and for the four-quarter-ahead results the weight on the AC-E17 forecast is much smaller. These results may be interpreted as indicating that the Fair model captures most of the information in the components of GNP.

V. Conclusion

The procedure used in this paper for examining models appears to be useful in comparing the different models. Using this procedure we have learned that the Fair model does very well relative to the other models. The Fair model cannot be dismissed as being based on the same information used in the other forecasts. The fact that the forecasts from the Fair model are significant shows that they are not collinear with the other forecasts and that the differences between the Fair model and the other models are meaningful.

We have also learned that information about components matters. The information about components of the kind incorporated in the AC models does help improve forecasts when compared with the VAR and AR models, but it is at best of modest benefit in improving the Fair model forecasts. In this sense the useful information in the Fair model not in the VAR or AR models includes information about components of GNP. We have also learned something about how to combine forecasts. While it appears that the VAR and AC forecasts do not contain a lot of information not in the Fair model forecasts for one-quarter-ahead forecasting, it may be that a combination of the forecasts from the Fair and VAR models is useful for four-quarter-ahead forecasting. While the AC model was dominated by the Fair model, it clearly contains information not in the VAR model. The VAR models seem to be losing useful information by ignoring the components of GNP and the GNP identity.

We should conclude with a warning about the interpretation of our results. The fact that one model does well or poorly for one sample period (in our case 1976 III–1986 II) does not necessarily mean that it will do well or poorly in future sample periods. The results could change if the structure of the economy is changing, which is, of course, true of any econometric result. In our case, however, the results could also change if the magnitudes of the forecast errors of the different models are changing at different rates through time. The errors could, for example, be changing at different rates because the data are providing different rates of improvement of the models' parameters.

APPENDIX

Components of AC Models

AC-6:

1. Personal consumption expenditures
2. Gross private fixed investment
3. Change in business inventories
4. Government purchases of goods and services
5. Exports
6. Imports

AC-13:

1. Personal consumption expenditures, durable goods
2. Personal consumption expenditures, nondurable goods
3. Personal consumption expenditures, services
4. Gross private fixed investment, nonresidential
5. Gross private fixed investment, residential
6. Change in business inventories, nonfarm
7. Change in business inventories, farm
8. Government purchases of goods, federal
9. Government purchases of goods, state and local
10. Government purchases of services, federal
11. Government purchases of services, state and local
12. Exports
13. Imports

AC-17:

1. Personal consumption expenditures, durable goods
2. Personal consumption expenditures, nondurable goods
3. Personal consumption expenditures, services
4. Gross private fixed investment, nonresidential, firm sector
5. Gross private fixed investment, nonresidential, financial sector
6. Gross private fixed investment, nonresidential, household sector
7. Gross private fixed investment, residential, firm sector
8. Gross private fixed investment, residential, financial sector
9. Gross private fixed investment, residential, household sector
10. Change in business inventories, firm sector
11. Change in business inventories, household sector
12. Government purchases of goods, federal
13. Government purchases of goods, state and local
14. Government purchases of services, federal
15. Government purchases of services, state and local
16. Exports

17. Imports

Note: See Ray C. Fair (1984) for the definitions of firm, financial, and household sectors. This breakdown is from the Flow of Funds Accounts.

AC-48:

Personal consumption expenditures, durable goods:
1. Motor vehicles and parts
2. Furniture and household equipment
3. Other

Personal consumption expenditures, nondurable goods:
4. Food
5. Clothing and shoes
6. Gasoline and oil
7. Fuel oil and coal
8. Other

Personal consumption expenditures, services:
9. Housing
10. Household operation, electricity and gas
11. Household operation, other
12. Transportation
13. Medical care
14. Other

Gross private fixed investment:
15. Nonresidential structures
16. Nonresidential producers' durable equipment
17. Residential

Change in business inventories:
18. Farm
19. Nonfarm, manufacturing, durable goods
20. Nonfarm, manufacturing, nondurable goods
21. Nonfarm, merchant wholesalers, durable goods
22. Nonfarm, merchant wholesalers, nondurable goods
23. Nonfarm, nonmerchant wholesalers, durable goods
24. Nonfarm, nonmerchant wholesalers, nondurable goods
25. Nonfarm, retail trade, durable goods
26. Nonfarm, retail trade, nondurable goods
27. Nonfarm, other, durable goods
28. Nonfarm, other, nondurable goods

Government purchases of goods and services, federal:
29. Durable goods
30. Nondurable goods
31. Services, compensation of employees, national defense, military
32. Services, compensation of employees, national defense, civilian
33. Services, compensation of employees, nondefense
34. Services, other services
35. Structures

Government purchases of goods and services, state and local:
36. Durable goods
37. Nondurable goods
38. Services, compensation of employees
39. Services, other services
40. Structures

Exports of goods and services:
41. Merchandise, durable goods
42. Merchandise, nondurable goods
43. Services, factor income
44. Services, other

Imports of goods and services:
45. Merchandise, durable goods
46. Merchandise, nondurable goods
47. Services, factor income
48. Services, other

TABLE A—PREDICTED VALUES AND ERRORS FOR THREE MODELS ONE-QUARTER-AHEAD FORECASTS

		FAIR		VAR4P3		AC-E17	
Quarter	Actual Change	Forecast Change	Error	Forecast Change	Error	Forecast Change	Error
1976.3	0.0042	0.0115	0.0073	0.0050	0.0008	0.0147	0.0105
4	0.0099	0.0104	0.0004	0.0043	−0.0056	0.0106	0.0007
1977.1	0.0136	0.0157	0.0021	0.0051	−0.0085	0.0221	0.0085
2	0.0159	0.0136	−0.0023	0.0060	−0.0099	0.0150	−0.0010
3	0.0200	0.0122	−0.0078	0.0061	−0.0139	0.0152	−0.0048
4	−0.0027	0.0116	0.0143	0.0066	0.0093	0.0124	0.0151
1978.1	0.0088	0.0105	0.0016	0.0038	−0.0051	0.0092	0.0004
2	0.0310	0.0103	−0.0207	0.0033	−0.0277	0.0019	−0.0291
3	0.0087	0.0112	0.0025	0.0054	−0.0033	0.0142	0.0056
4	0.0123	0.0066	−0.0057	0.0042	−0.0081	0.0110	−0.0013
1979.1	0.0000	0.0049	0.0048	0.0032	0.0031	0.0043	0.0042
2	−0.0010	0.0102	0.0112	0.0010	0.0020	0.0062	0.0073
3	0.0091	0.0087	−0.0003	0.0001	−0.0090	0.0017	−0.0073
4	−0.0019	0.0076	0.0095	0.0013	0.0032	0.0074	0.0093
1980.1	0.0099	0.0009	−0.0090	0.0002	−0.0097	0.0059	−0.0040
2	−0.0238	0.0055	0.0293	0.0005	0.0243	0.0054	0.0291
3	0.0006	0.0133	0.0127	−0.0005	−0.0010	0.0008	0.0002
4	0.0127	0.0188	0.0062	0.0017	−0.0109	0.0051	−0.0076

TABLE A—CONTINUED

		FAIR		VAR4P3		AC-E17	
Quarter	Actual Change	Forecast Change	Error	Forecast Change	Error	Forecast Change	Error
1981.1	0.0191	0.0059	−0.0132	0.0020	−0.0171	0.0101	−0.0090
2	−0.0033	0.0021	0.0054	0.0036	0.0069	0.0032	0.0065
3	0.0045	0.0030	−0.0014	0.0023	−0.0022	0.0084	0.0039
4	−0.0140	−0.0031	0.0109	0.0022	0.0163	−0.0005	0.0135
1982.1	−0.0153	−0.0017	0.0136	0.0026	0.0179	0.0024	0.0177
2	0.0030	0.0052	0.0022	0.0021	−0.0009	−0.0008	−0.0037
3	−0.0080	0.0044	0.0124	0.0048	0.0128	0.0049	0.0129
4	0.0016	0.0034	0.0019	0.0060	0.0044	0.0018	0.0002
1983.1	0.0085	0.0229	0.0143	0.0081	−0.0004	0.0200	0.0115
2	0.0223	0.0198	−0.0025	0.0099	−0.0124	0.0158	−0.0065
3	0.0147	0.0190	0.0044	0.0125	−0.0021	0.0136	−0.0011
4	0.0177	0.0125	−0.0051	0.0119	−0.0058	0.0191	0.0014
1984.1	0.0233	0.0154	−0.0079	0.0112	−0.0121	0.0146	−0.0087
2	0.0123	0.0078	−0.0044	0.0113	−0.0010	0.0078	−0.0044
3	0.0057	0.0072	0.0015	0.0100	0.0042	0.0073	0.0016
4	0.0037	0.0070	0.0033	0.0079	0.0043	0.0076	0.0039
1985.1	0.0076	0.0128	0.0053	0.0079	0.0003	0.0032	−0.0044
2	0.0058	0.0131	0.0073	0.0084	0.0026	0.0072	0.0014
3	0.0101	0.0125	0.0024	0.0093	−0.0008	0.0044	−0.0056
4	0.0052	0.0145	0.0094	0.0107	0.0055	0.0094	0.0042
1986.1	0.0092	0.0179	0.0087	0.0103	0.0011	0.0074	−0.0018
2	0.0027	0.0139	0.0112	0.0107	0.0080	0.0031	0.0004

Four-Quarter-Ahead Forecasts

		FAIR		VAR4P3		AC-E17	
Quarter	Actual Change	Forecast Change	Error	Forecast Change	Error	Forecast Change	Error
1977.2	0.0436	0.0559	0.0122	0.0163	−0.0274	0.0524	0.0088
3	0.0595	0.0569	−0.0025	0.0161	−0.0434	0.0548	−0.0047
4	0.0468	0.0693	0.0224	0.0189	−0.0279	0.0695	0.0226
1978.1	0.0421	0.0539	0.0118	0.0205	−0.0216	0.0482	0.0061
2	0.0571	0.0492	−0.0079	0.0192	−0.0379	0.0429	−0.0142
3	0.0458	0.0467	0.0010	0.0196	−0.0261	0.0321	−0.0136
4	0.0608	0.0437	−0.0171	0.0131	−0.0477	0.0218	−0.0389
1979.1	0.0520	0.0402	−0.0117	0.0122	−0.0398	0.0234	−0.0286
2	0.0200	0.0364	0.0165	0.0156	−0.0043	0.0293	0.0094
3	0.0204	0.0375	0.0171	0.0125	−0.0078	0.0222	0.0018
4	0.0062	0.0267	0.0205	0.0091	0.0029	0.0181	0.0120
1980.1	0.0161	0.0337	0.0177	0.0042	−0.0118	0.0145	−0.0016
2	−0.0067	0.0245	0.0311	0.0028	0.0095	0.0096	0.0163
3	−0.0152	0.0213	0.0365	0.0066	0.0218	0.0208	0.0359
4	−0.0006	0.0158	0.0164	0.0017	0.0023	0.0276	0.0282
1981.1	0.0086	0.0151	0.0066	0.0015	−0.0071	0.0322	0.0237
2	0.0290	0.0315	0.0025	0.0084	−0.0206	0.0058	−0.0233
3	0.0329	0.0494	0.0165	0.0142	−0.0187	0.0263	−0.0066
4	0.0062	0.0159	0.0097	0.0075	0.0013	0.0335	0.0273
1982.1	−0.0282	0.0113	0.0395	0.0105	0.0387	0.0275	0.0557
2	−0.0219	0.0081	0.0300	0.0092	0.0311	0.0241	0.0460
3	−0.0344	−0.0154	0.0190	0.0102	0.0446	0.0104	0.0448
4	−0.0188	0.0040	0.0227	0.0177	0.0364	0.0223	0.0411
1983.1	0.0050	0.0155	0.0105	0.0169	0.0119	0.0270	0.0220
2	0.0244	0.0164	−0.0080	0.0249	0.0006	0.0317	0.0073
3	0.0471	0.0227	−0.0244	0.0296	−0.0174	0.0273	−0.0197
4	0.0632	0.0788	0.0156	0.0368	−0.0264	0.0563	−0.0069
1984.1	0.0779	0.0680	−0.0099	0.0394	−0.0385	0.0516	−0.0264
2	0.0679	0.0621	−0.0058	0.0447	−0.0231	0.0581	−0.0098
3	0.0590	0.0486	−0.0104	0.0407	−0.0182	0.0583	−0.0006
4	0.0450	0.0501	0.0052	0.0374	−0.0076	0.0385	−0.0064
1985.1	0.0292	0.0309	0.0017	0.0364	0.0071	0.0205	−0.0087
2	0.0228	0.0230	0.0002	0.0338	0.0110	0.0191	−0.0037
3	0.0271	0.0246	−0.0025	0.0298	0.0026	0.0106	−0.0165
4	0.0286	0.0416	0.0130	0.0327	0.0041	0.0105	−0.0181
1986.1	0.0303	0.0407	0.0104	0.0348	0.0045	0.0147	−0.0156
2	0.0272	0.0415	0.0143	0.0382	0.0110	0.0134	−0.0137

REFERENCES

Andrews, Donald W. K., "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," Cowles Foundation Discussion Paper No. 877R, July 1989.

Chong, Yock Y. and Hendry, David F., "Econometric Evaluation of Linear Macro-Econometric Models," *Review of Economic Studies*, August 1986, *53*, 671–90.

Cooper, J. Phillip and Nelson, Charles R., "The *Ex-Ante* Prediction Performance of the St. Louis and FRB-MIT-Penn Econometric Models and Some Results on Composite Predictions," *Journal of Money, Credit, and Banking*, February 1975, *7*, 1–32.

Cumby, Robert E., Huizinga, John and Obstfeld, Maurice, "Two-Step Two Stage Least Squares in Models with Rational Expectations," *Journal of Econometrics*, April 1983, *21*, 333–55.

Davidson, Russell, and MacKinnon, James G., "Several Tests of Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, May 1981, *40*, 781–93.

Fair, Ray C., *A Model of Macroeconomic Activity, Vol. 2, The Empirical Model*, Cambridge MA: Ballinger Publishing, 1976.

______, "The Sensitivity of Fiscal Policy Effects to Assumptions About the Behavior of the Federal Reserve," *Econometrica*, September 1978, *46*, 1165–79.

______, *Specification, Estimation, and Analysis of Macroeconometric Models*, Cambridge, MA: Harvard University Press, 1984.

Granger, Clive W. J. and Newbold, Paul, *Forecasting Economic Time Series* 2nd ed., New York: Academic Press, 1986.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, July 1982, *50*, 1029–54.

Hendry, David F. and Richard, Jean-Francois, "On the Formulation of Empirical Models in Dynamic Economics," *Journal of Econometrics*, October 1982, *20*, 3–33.

Litterman, Robert B., "Techniques of Forecasting Using Vector Autoregression," Federal Reserve Bank of Minneapolis Working Paper No. 115, November 1979.

McNees, Stephen K., "The Methodology of Macroeconometric Model Comparisons," in J. Kmenta and J. B. Ramsey, eds., *Large Scale Macroeconometric Models*, Amsterdam: North-Holland, 1981, pp. 397–442.

Mizon, Grayham and Richard, Jean-Francois, "The Encompassing Principle and Its Application to Testing Non-Nested Hypotheses," *Econometrica*, May 1986, *54*, 657–78.

Murphy, Kevin M. and Topel, Robert H., "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics*, October 1985, *3*, 370–79.

Nelson, Charles R., "The Prediction Performance of the FRB-MIT-Penn Model of the U.S. Economy," *American Economic Review*, December 1972, *62*, 902–17.

Sims, Christopher A., "Macroeconomics and Reality," *Econometrica*, January 1980, *48*, 1–48.

______, "Economic Forecasts from a Vector Autoregression, mimeo., serial.

White, Halbert, "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48*, 817–38.

White, Halbert and Domowitz, Ian, "Nonlinear Regression with Dependent Observations," *Econometrica*, January 1984, *52*, 143–61.

Money, Output, and the Nominal National Debt

By Bruce Champ and Scott Freeman*

This paper presents a model of finitely lived rational agents in which unanticipated innovations in the stock of fiat money affect real variables. An unanticipated inflation reduces the real value of the nominally denominated national debt, thereby reducing the crowding-out of capital and/or the tax burden. Both effects stimulate increased investment in capital, which leads to an increase in real output and wages in the following periods. In contrast with price-surprise models, these real effects occur even if the monetary innovation is instantly and perfectly observed by agents. (JEL 311)

The possibility that innovations in the monetary base are correlated with innovations in important real variables raises difficult yet fascinating questions of theory and policy for macroeconomists committed to the rigor imposed by the assumption of choice-theoretic models of rational agents. How might the printing of intrinsically useless pieces of paper be linked to the real output of the nation's workers and machines? Might there be a way to systematically exploit real effects of the monetary base?

These questions were first addressed by a choice-theoretic model with the publication of "Expectations and the Neutrality of Money" (Robert Lucas, 1972). Following the lead of this greatly influential paper, recent studies of the effects of monetary forces on real economic aggregates have often focused on the reaction of agents who have misinterpreted nominal price shocks as shocks to the real demand for their output. While the Lucas model can account for a difference in the correlations of anticipated and unanticipated inflation with real variables, the assumption in price-surprise/information-confusion models that agents lack information on the current money stock and price level is inconsistent with the ready availability of current information on prices and monetary aggregates, most of which are reported with delays of only a week or a month. We are led, therefore, to look for alternative explanations for the transmission of nominal monetary innovations to real variables.

Another means by which monetary policy can conceivably affect real variables is through the real revenue raised by the printing of fiat money. If this seigniorage is applied to the purchase of government bonds in an economy where government bonds are net wealth, money creation will reduce the real interest rate, stimulating investment and output. The use of seigniorage as a means of reducing the burden of the national debt is explored in Lloyd Metzler (1951), Preston Miller (1981), Thomas Sargent and Neil Wallace (1981), Wallace (1984), and Douglas Waldo (1985). However, the revenue from the tax on the monetary base is a small fraction of the size of total investment, output, and government revenue in most industrialized countries.[1] One wonders how this small tax can account for major changes in output.

A revenue effect of inflation that has been a recent focus of attention is the change in the real value of the outstanding nominally denominated government debt (see Jeremy Siegel, 1979, and Robert Eisner and Paul Pieper, 1984). Given that the U.S. federal

*Respectively, Department of Economics, University of Western Ontario, London, Ontario N6A 5C2, and Department of Economics, University of California, Santa Barbara, CA 93106.

[1] See Barro (1982) or Fischer (1982).

debt is roughly ten times the size of the monetary base, the revenue effect of an inflation-induced change in the real value of the national debt will exceed the revenue from the taxation of the monetary base through seigniorage.[2] With this in mind, we propose in this paper a model in which the primary effect of monetary policy is the change in the real value of the nominally denominated national debt. In this model unanticipated inflation reduces the real value of the national debt in much the same way as a default. The effect of unanticipated inflation on the real national debt has captured the attention of those (such as Guillermo Calvo, 1978, 1988; Lucas and Nancy Stokey, 1983; and Robert Barro and David Gordon, 1983) examining the time consistency problem of government finance. Our work follows their lead with one important difference—agents in our model have finite planning horizons (lives). The assumption of finite planning horizons implies that the change in the real debt will affect the distribution of personal wealth among agents at different stages in their life cycle and thus consumption, capital accumulation, output, and other macroeconomic variables.

The model is a version of Peter Diamond's (1965) overlapping generations model in which fiat money is held to satisfy legally imposed reserve requirements and where the government debt is denominated in units of fiat money. Agents may save for old age with either capital or government bonds, which are viewed as net wealth because of the finite planning horizon of each agent. Money is assumed to offer no transaction services in order to demonstrate that the real effects of monetary policy follow from its fiscal effects, not its effects on transaction costs or the mechanics of exchange.

This model shares some features with other models. As in the IS-LM model, for example, monetary policy works only to the extent that it lowers the real interest rate, stimulating investment and output. As in disequilibrium models, monetary policy has an effect because a critical value (the national debt in our model, prices in disequilibrium models) is fixed in nominal terms. However, unlike the static IS-LM model or fixed-price disequilibrium models, the model of this paper is constructed from a fully explicit general equilibrium model of utility maximizing rational agents, with no *ad hoc* restrictions on the behavior of agents or prices. As a result it makes clear distinctions between the effects of changes in the moments of the monetary policy process and the effects of different realizations of a given monetary policy process. In this way the model is consistent with the classical belief in the "long run" neutrality of monetary policy.

I. The Model

A. *The Environment*

The model is an adaptation of Diamond's (1965) version of Samuelson's (1958) overlapping generations model. In each period $t \geq 1$, N_t two-period lived agents are born. Let $N_t = nN_{t-1}$ for each period t, where n is a positive constant. An agent born at t maximizes the expected value of the twice continuously differentiable utility function $U(c_{1t}) + E_t c_{2t}$, where c_{it} denotes the agent's consumption of the economy's sole consumption good in the ith period of life. The function $U(\cdot)$ is strictly concave and strictly increasing, with $\lim_{c \to 0} U'(c) = \infty$.

Each agent born in period $t \geq 1$ is endowed with one unit of labor when young (and nothing when old), which is supplied inelastically to the labor market. There also exist N members of generation 0 (to be called the initial old) who are alive only in period 1. The utility of each initial old agent is a strictly increasing function of personal consumption in period 1.

In any period, one unit of capital can be created from every unit of the consumption good that is not consumed. We adopt the constant-returns-to-scale production function of Diamond (1965), where production of the good at t is a function, $F(K_{t-1}, L_t)$,

[2] The Federal Reserve Bulletin of October 1988 puts the federal debt at $2,493 billion (Table 1.40, page A-30) and the monetary base at $263 billion (Table 1.20, page A-12).

of capital created at $t-1, K_{t-1}$, and labor supplied at t, L_t. The function $F(\cdot,\cdot)$ is twice continuously differentiable, concave, and increasing in each argument. Let k_t represent the units of capital created by an individual in any period t. The assumption of constant returns to scale allows us to define production at $t+1$ as a function of the capital created per young person at t: $f(k_t) \equiv F(k_t, n)$ units of the consumption good in period $t+1$. It follows that the function $f(k_t)$ is strictly increasing, concave, and twice continuously differentiable. Capital created at t completely disintegrates at the end of period $t+1$.

The initial old are endowed with M_0 units of fiat money, which consists of unbacked, intrinsically useless pieces of paper costlessly produced by the government. The stock of fiat money changes according to the rule $M_t = z_t M_{t-1}$, where M_t denotes the money stock at time t after the change in the money stock has taken place. The change in the money stock at t is accomplished through lump-sum subsidies of $a_t \equiv (z_t - 1)M_{t-1}/N_{t-1}$ units of fiat money to each old person. The gross rate of change of the fiat money stock is $z_t = z_t^*/(1-\varepsilon_t)$, where z_t^* is a positive random variable and ε_t is a zero-mean, serially uncorrelated random variable with $\varepsilon_t < 1\ \forall t$. The realizations of z_t^* are known by all with complete certainty at $t-1$, but the realizations of ε_t are not known until t. We will therefore label z_t^* as the "anticipated" rate of fiat money creation at t, while ε_t will represent the unanticipated innovation at t in the money creation process. Each young agent is required by law to hold fiat money worth at least γ units of the consumption good. Let p_t denote the price of a good in units of fiat money.

In each period t the government uses up g_t goods per young agent in some project that has no direct effect on the utility or consumption of agents. In each period t the government is also assumed to have an outstanding nominally denominated debt worth B_t units of fiat money. This government debt has a one-period maturity and pays the nominal gross rate of return R_t at $t+1$. In the initial period the initial old own a total of B_0 units of nominal government debt and so are owed $R_0 B_0$ units of fiat money in period 1.

To finance its expenditures and debt obligations, the government collects a lump-sum tax of τ_t units of the consumption good from each young person in each period t. Note that seigniorage revenue is assumed to be used only to finance subsidies so that it will not finance any part of government expenditures, interest payments, or debt retirement. This is assumed so that all the revenue effects of the expansion of the fiat money stock in this model will come solely from the effects of inflation on the real value of national debt. The government budget constraints in period t can now be written as:

$$N_t p_t g_t + R_{t-1}B_{t-1} = B_t + p_t N_t \tau_t, \tag{1}$$

$$M_t = M_{t-1} + N_{t-1}a_t. \tag{2}$$

B. *Equilibrium Conditions*

Define a rational expectations competitive equilibrium in this economy as values of the endogenous variables (listed below) such that (1) firms choose capital and labor to maximize profits while taking factor prices as given, (2) agents choose their assets to maximize expected utility taking the probability distributions of rates of return and the price level as given, (3) the probability distributions of rates of return and the price level anticipated by agents when solving their maximization problems are the actual equilibrium probability distributions, and (4) markets clear.

Perfectly competitive markets in capital and labor will ensure that x_{t-1}, the marginal gross real return in period t to a unit of capital created at $t-1$, is equal to its marginal product, or

$$x_{t-1} = F_K(k_{t-1}, n) \equiv f'(k_{t-1}), \tag{3}$$

and that the real wage, the return to a unit of labor at t (w_t), is its marginal product,

$$\begin{aligned} w_t &= F_L(k_{t-1}, n) \\ &\equiv [f(k_{t-1}) - f'(k_{t-1})k_{t-1}]/n. \end{aligned} \tag{4}$$

The maximization problem of an individual agent born at t can now be expressed as the choice of personal holdings of capital (k_t), bonds, (b_t), and fiat money (m_t) to maximize expected utility $U(c_{1t}) + E_t c_{2t}$ subject to his budget constraints:

$$(5) \quad c_{1t} = w_t - \tau_t - k_t - b_t/p_t - m_t/p_t$$

$$(6) \quad c_{2t} = x_t k_t + R_t b_t/p_{t+1} + m_t/p_{t+1} + a_{t+1}/p_{t+1}$$

$$(7) \quad m_t/p_t \geq \gamma.$$

The following market clearing constraints

$$(8) \quad B_t = N_t b_t,$$

$$(9) \quad M_t = N_t m_t,$$

and the government budget constraints (1) and (2) must also be met in equilibrium.

We will confine our attention to equilibria in which the legal requirement to hold fiat money is binding $(m_t/p_t = \gamma \ \forall \ t)$, equilibria in which the rate of return on capital exceeds the expected rate of return on fiat money $[f'(k) > n]$ so that fiat money is not viewed as a substitute for capital. Notice first that equations (2) and (9) imply that $m_{t-1} + a_t = nm_t$. When the legal restriction (7) holds with equality, the problem of an agent born at t is simplified to the choice of real capital (k_t) and nominal bonds (b_t) to maximize expected personal utility. The budget constraints (5) and (6) can now be written as:

$$(10) \quad c_{1t} = w_t - \tau_t - k_t - b_t/p_t - \gamma$$

$$(11) \quad c_{2t} = x_t k_t + R_t b_t/p_{t+1} + n\gamma.$$

The resulting first-order conditions for a maximum can be written as:

$$(12) \quad U'[w_t - \tau_t - k_t - b_t/p_t - \gamma] = f'(k_t)$$

and

$$(13) \quad x_t = f'(k_t) = E_t[R_t p_t/p_{t-1}].$$

The first (12) describes an agent's saving behavior, and the second (13) represents an arbitrage condition guaranteeing the equality of the expected real rates of return of capital and bonds.

We can use (13) to find an expression for the nominal interest rate R_t as a function of capital and expected inflation. First, we will need an expression for the inverse of the inflation rate, the rate of return on fiat money. From (6) and (8),

$$(14) \quad p_t = M_t/N_t\gamma$$

$$(15) \quad \Rightarrow p_t/p_{t+1} = M_t N_{t+1}\gamma/M_{t+1}N_t\gamma = n/z_{t+1} = n(1-\varepsilon_{t+1})/z^*_{t+1}.$$

Then $E_t[R_t p_t/p_{t+1}] = R_t E_t[n(1-\varepsilon_{t+1})/z_{t+1}{}^*] = R_t n/z^*_{t+1}$ and from (13):

$$(16) \quad R_t = f'(k_t) z^*_{t+1}/n.$$

The nominal interest rate on bonds is therefore a function solely of the expected rate of money creation, the rate of growth of the economy (n), and the marginal product of capital. Notice that $R_t > 1$ if, as assumed, $f'(k_t) > n/z^*_{t+1}$.

The Real Effects of Inflation. We are now ready to examine the effects of anticipated and unanticipated inflation on the capital stock, the real interest rate, and output. The central real variable in this economy is the capital stock. The real interest rate at t is the marginal product of capital, $f'(k_t)$, a decreasing function of capital. Real output and the real wage at $t+1$ are increasing functions of capital created at t. The effects of anticipated and unanticipated inflation on the capital stock are easy to determine and are presented in Propositions 1 and 2.

PROPOSITION 1: *The real values of capital, output, the wage, and the interest rate are independent of z^*_{t+1}, the anticipated rate of money creation.*

PROPOSITION 2: *A positive realization of ε_t (an unanticipated positive innovation in the fiat money stock) is associated with an increase in the current capital, a decrease in the current real and nominal interest rates, and an increase of real output and the real wage one period later.*

We will prove these propositions by finding capital as a reduced form function of z and ε_t. From the government budget constraint (1), we can find the following expression for real government debt per young agent at t:

$$(17)\quad B_t/p_tN_t = R_{t-1}B_{t-1}/p_tN_t + g_t - \tau_t.$$

Substitution of this expression for b_t/p_t into the first-order condition describing desired saving (12) yields

$$(18)\quad U'[w_t - R_{t-1}B_{t-1}/p_tN_t - g_t - k_t - \gamma] = f'(k_t).$$

Define $d_t \equiv R_{t-1}B_{t-1}/p_tN_t$ as the real burden of the national debt passed on to each young agent at t, that is, the current value of the last period's debt (including interest).

LEMMA 1: *Capital at t per young agent (k_t) is a decreasing function of the burden of the national debt at t per young agent (d_t) and of government expenditures at t per young person (g_t) and an increasing function of the real wage at t (w_t).*

Lemma 1 can be quickly proven by applying the implicit function theorem to (18).

It remains to determine the effect of z^*_{t+1} and ε_t on the current value of last period's debt.

$$\begin{aligned}(19)\quad d_t &= R_{t-1}B_{t-1}/p_tN_t\\ &= R_{t-1}[B_{t-1}/p_{t-1}N_{t-1}](p_{t-1}/p_tn)\\ &= R_{t-1}[B_{t-1}/p_{t-1}N_{t-1}]\\ &\quad\times(1-\varepsilon_t)/z^*_t \qquad \text{from (14)}\\ &= f'(k_{t-1})(z^*_t/n)[B_{t-1}/p_{t-1}N_{t-1}]\\ &\quad\times(1-\varepsilon_t)/z^*_t \qquad \text{from (15)}\\ &= [f'(k_{t-1})/n][B_{t-1}/p_{t-1}N_{t-1}]\\ &\quad\times(1-\varepsilon_t).\end{aligned}$$

Notice from (18) that the current value of last period's debt is unaffected by the rate of anticipated money creation (z^*_t) but is a function of ε_t. A positive value of ε_t represents an unanticipated positive innovation in the fiat money stock and, hence, a negative innovation in the rate of return of fiat money. Since the debt is denominated in units of fiat money, an unanticipated inflation acts like a partial default and lowers the value of the outstanding national debt. Since the capital stock at t is a decreasing function of government debt at t (see Lemma 1), such an unanticipated inflation will increase the formation of new capital, reducing the marginal product of capital created at t (the expected real interest rate of assets paying a return at $t+1$) and increasing output at $t+1$. (Output responds to unanticipated inflation with a lag because of the assumption that it takes one period for new capital to augment production. A more general assumption would be that new capital contributes to both current and future output, resulting in both a contemporaneous and lagged response of output to unanticipated inflation. This case is presented in the appendix.)

Unanticipated inflation redistributes wealth from the current old to future generations. The current young react by increasing investment regardless of the method of financing the burden of past debt. If the government responds to the smaller burden of past debt with a reduction in new debt, the young replace bonds with capital in their portfolios. If, on the other hand, it responds with a reduction in taxes, the young are wealthier and thus desire to save more (through increased capital) for their old age.

The nonneutrality of government debt is crucial to the implications of the model. If all agents had infinite lives, changes in government debt would have no effect on current consumption or the capital stock—the wealth increment from any increase in real government debt would be exactly offset by the expected burden of future taxes to service the debt. To pay the expected tax burden, an infinitely lived agent would increase his savings by exactly the amount of the increase in the real debt, preventing the increased national debt from crowding out capital. As Barro (1974) has argued, agents who value the utility of their offspring and actively leave them bequests will react to changes in the national debt in the same way as infinitely lived agents.

The model's implication that the resulting change in the capital stock is of the same magnitude whether the burden of the past debt is financed by current taxes or additional debt follows from the constant marginal utility of consumption by the old. It is a great convenience because we are able to describe the current effects of the burden of debt in an equation as simple as (18) without having to refer to the tax policy through which the debt is to be financed in future periods.

In this model the means of financing the added burden of past debt at t matter only to the values of the capital stock after t. If the added burden of past debt is financed through new bonds, the new debt will continue to reduce capital in the periods after t; if the added burden of past debt is financed through taxes, the reduction of capital takes place only at t.

Independent of the means by which the government finances the burden of past debt, serially uncorrelated, unanticipated innovations in the money stock lead to serially correlated innovations in output. Unanticipated inflation at t, by increasing capital formed at t, increases output and wages at $t+1$, which in turn increase capital at $t+1$ and so output at $t+2$.

Anticipated inflation has no effect on real activity through changes in the real national debt because the anticipated real return on nominal bonds is tied through arbitrage to the real return on capital. This implication follows from two special features of the model—the demand for money is fixed and additions to the fiat money stock are distributed back to agents. If the demand for fiat money were not fixed, real money balances would respond to changes in anticipated inflation with the effects on real capital and output suggested by James Tobin (1965). If the seigniorage from the expansion of the fiat money stock were not returned to agents but used to help finance government expenditures or to retire the debt, an anticipated expansion of the fiat money stock would have the same qualitative effects as an unanticipated expansion. The effects would be of smaller magnitude because an anticipated inflation taxes only fiat money balances, while an unanticipated inflation taxes both fiat money balances and nominally denominated national debt.

Unanticipated monetary innovations have real effects in this economy even if the realization of the money growth process is instantly and perfectly observed by all. In this way the model differs from that of Lucas with its assumption that agents confuse real and nominal shocks because they cannot observe the behavior of the money stock. Since information on monetary aggregates is available weekly, the lack of information about the current money supply is an implausible explanation for output fluctuations that last more than a week. In contrast, in the model of this paper, a monetary innovation may be known with no delay yet will nonetheless affect output through a devaluation of the outstanding nominal stock of government debt. It affects output and the real value of the debt because bonds are purchased at fixed nominal rates of interest before the realization of the money stock is known.

II. Concluding Remarks

We have proposed in this paper a means through which changes in the stock of fiat money may affect such real variables as investment, the real interest rate, and output. It has an effect when unanticipated inflationary innovations reduce the real value of the national debt. Of course, this effect has been considered before, in the recent work of Lucas and Stokey (1983) and Barro and Gordon (1983) and in work as early as that of Arthur Pigou (1943) and Irving Fisher (1933). The model of this paper can be distinguished from these efforts by its explicitly modeled life-cycle properties. In this model inflation-induced reductions in the national debt increase output through increased capital holding in one of two ways. To the extent that the reduction of the value of the old national debt reduces current taxes it redistributes wealth from the owners of the bonds, the current old, with no propensity to save, to the young, who at this point in their life cycle are ready to invest. To the extent that it reduces the new level of government debt, it reduces the crowding out of capital in agent portfolios. Neither effect is an implication of models of infinitely lived representative agents, where reductions in the burden of the debt matter only through the reduction of the distortionary taxes needed to finance the debt. The potential importance of the redistribution of wealth through inflation was recognized by economists as early as Fisher (1933),[3] but not in explicit models (by today's standards) of the means through which the redistribution affected output. The model presented here, with its explicit differentiation of the investment behavior of the owners of mature bonds (the current old) and that of those about to invest in capital and bonds (the current young), is offered as a structure in which we may examine the macroeconomic effects of the redistribution of wealth by unanticipated inflation.

The model is also offered as an alternative to Lucas's price-surprise explanation of the observed difference in the effects of anticipated and unanticipated inflation, an alternative that does not require the obviously counterfactual assumption that the current money stock is not observable. With its assumption of rational expectations, this model does share the policy implications of the Lucas model inasmuch as inflationary innovations cannot be systematically used to stimulate output.

APPENDIX

The model presented in the main body of the paper predicts that output will respond to unanticipated innovations in money with a lag. The lag comes from the model's adoption of Diamond's production technology, which assumes that investment produces output after a one-period delay. Given that one period represents half of a lifetime in this model, the assumed delay in production seems very long relative to actual production lags and the observed response of output to money. We may capture a more contemporaneous response of output to unanticipated monetary innovations in a version of the model in which capital produces goods in both the period of its creation and in the following period. In particular, consider a version of the model in which the creation at t of K_t units of capital produces $h_t K_t$ goods in period t in addition to the previously assumed $F(K_t, L_{t+1})$ goods in period $t+1$.

The budget constraints of a young agent born at t corresponding to (10) and (11) can now be written as:

$$c_{1t} = w_t + h_t k_t - \tau_t - k_t - b_t/p_t - \gamma, \tag{20}$$

$$c_{2t} = x_t k_t + R_t b_t/p_{t+1} + n\gamma. \tag{21}$$

The resulting first-order conditions for a maximum can be written as:

$$\begin{aligned} U'[w_t - \tau_t - k_t - b_t/p_t - \gamma] &= x_t/(1-h_t) \\ &= f'(k_t)/(1-h_t) \\ &= E_t[R_t p_t/p_{t+1}]. \end{aligned} \tag{22}$$

When combined with the government budget constraint we can define capital per young person, k_t, as a function of the real burden of past debt, d_t, as we did in (18):

$$U'[w_t - d_t - g_t - k_t - \gamma](1-h_t) = f'(k_t). \tag{23}$$

As before, we find that k_t is a decreasing function of real burden of past debt, d_t.

In this economy, aggregate real output in period t is the sum of production from previously and currently created capital:

$$N_{t-1} f(k_{t-1}) + N_t h_t k_t. \tag{24}$$

Therefore, unanticipated inflation at t, by reducing the real burden of past debt, will stimulate capital creation at t, increasing output at both t and $t+1$. By reducing the assumed lag between investment and output in this way, we find both a contemporaneous and a lagged response of output to monetary innovations.[4]

[3] A clear summary and comparison of these views may be found in Tobin (1980).

[4] We might also have found arbitrarily more "realistic" timing of the response of output to money by assuming that agents live $T > 2$ period lives. As long as T is finite, the real burden of past debt will affect capital and then output. The lag between the two can then be set to any length that appropriately represents the time that it takes for new capital to start to produce. However, a model with more general life spans and production delays immediately loses a great deal of tractability, which influenced us to remain within the friendly confines of the simple model of two-period lives.

To maintain the tractability of the model [the closed form solution of (23)], we have assumed that the contemporaneous production, h_t, from newly created capital does not require a labor input. If it did, unanticipated inflation would stimulate not only current output but the current real wage as well, as the increased capital raises the marginal product of labor.

REFERENCES

Barro, Robert J., "Are Government Bonds Net Wealth?" *Journal of Political Economy*, November/December 1974, *91*, 1095–1117.

______, "Measuring the Fed's Revenue from Money Creation," *Economic Letters*, 1982, *3–4*, 327–32.

______ **and Gordon, David B.**, "A Positive Theory of Monetary Policy in a Natural Rate Model," *Journal of Political Economy*, August 1983, *91*, 590–610.

Calvo, Guillermo A., "On the Time Consistency of Optimal Policy in a Monetary Economy," *Econometrica*, November 1978, *46*, 1411–28.

______, "Servicing the Public Debt: The Role of Expectations," *American Economic Review*, September 1988, *78*, 647–61.

Cass, David and Yaari, Menahem, "A Reexamination of the Pure Consumption Loans Model," *Journal of Political Economy*, August 1966, *74*, 353–67.

Diamond, Peter A., "National Debt in a Neoclassical Growth Model," *American Economic Review*, December 1965, *55*, 1126–50.

Eisner, Robert and Pieper, Paul J., "A New View of the Federal Debt and Budget Deficits," *American Economic Review*, March 1984, *74*, 11–20.

Fisher, Irving, "The Debt Deflation Theory of Great Depressions," *Econometrica*, October 1933, *1*, 337–50.

Fischer, Stanley, "Seigniorage and the Case for a National Money," *Journal of Political Economy*, April 1982, *90*, 295–313.

Lucas, Robert E., Jr., "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April 1972, *4*, 103–24.

______ **and Stokey, Nancy**, "Optimal Fiscal and Monetary Policy in an Economy without Capital," *Journal of Monetary Economics*, July 1988, *22*, 55–93.

Metzler, Lloyd, A., "Wealth, Saving, and the Rate of Interest," *Journal of Political Economy*, April 1951, *59*, 93–116.

Miller, Preston, J., "Fiscal Policy in a Monetarist Model," Federal Reserve Bank of Minneapolis Staff Report No. 67, 1981.

Pigou, A. C., "The Classical Stationary State," *Economic Journal*, December 1943, *53*, 343–51.

Samuelson, Paul A., "An Exact Consumption-Loan Model of Interest with or Without the Social Contrivance of Money," *Journal of Political Economy*, December 1958, *66*, 467–82.

Sargent, Thomas and Wallace, Neil, "Some Unpleasant Monetarist Arithmetic," Federal Reserve Bank of Minneapolis *Quarterly Review*, Fall 1981, 1–17.

Siegel, Jeremy J., "Inflation-Induced Distortions in Government and Private Saving Statistics," *Review of Economics and Statistics*, February 1979, *61*, 83–90.

Tobin, James, "Money and Economic Growth," *Econometrica*, October 1965, *33*, 671–84.

______, *Asset Accumulation and Economic Activity*, Chicago: University of Chicago Press, 1980.

Wallace, Neil, "Some of the Choices for Monetary Policy," Federal Reserve Bank of Minneapolis *Quarterly Review*, Winter 1984, 15–24.

Waldo, Douglas G., "Open Market Operations in an Overlapping Generations Model," *Journal of Political Economy*, December 1985, *93*, 1243–57.

Mean Reversion in Equilibrium Asset Prices

By Stephen G. Cecchetti, Pok-sang Lam, and Nelson C. Mark*

This paper demonstrates that negative serial correlation in long horizon stock returns is consistent with an equilibrium model of asset pricing. When investors display only a moderate desire to smooth their consumption, commonly used measures of mean reversion in stock prices calculated from historical returns data nearly always lie within a 60 percent confidence interval of the median of the Monte Carlo distributions implied by our equilibrium model. From this evidence, we conclude that the degree of serial correlation in the data could plausibly have been generated by our model. (JEL 313)

Recent research into the behavior of the stock market reports evidence that returns are negatively serially correlated. James Poterba and Lawrence Summers (1988) find that variance ratio tests reject the hypothesis that stock prices follow a random walk, and Eugene Fama and Kenneth French (1988) show that there is significant negative autocorrelation in long-horizon returns.[1] It is well known (see Stephen Leroy, 1973; Robert Lucas, 1978; and Ronald Michener, 1982) that serial correlation of returns does not in itself imply a violation of market efficiency.[2] Nevertheless, there is a tendency to conclude that evidence of mean reversion in stock prices constitutes a rejection of equilibrium models of rational asset pricing. Fama and French suggest this interpretation as a logical possibility, while Poterba and Summers argue that the serial correlation in returns should be attributed to "price fads." In this paper we demonstrate that the serial correlation in returns that is computed from stock market data is consistent with an equilibrium model of asset pricing.

Our approach is to combine the methods of model calibration and statistical inference to critically evaluate the conclusions that can be drawn from the available data. We begin by specifying both an economic model of asset pricing and a stochastic model for the exogenous forcing process driving the economic fundamentals. The forcing process is then calibrated to actual data from the U.S. economy over a long historical period. From this structure, we compute the Monte Carlo distributions of the statistics that previous investigators have used, under the null hypothesis that our rational equilibrium model is true. Finally, we can state the likelihood that those statistics, computed with historical returns, were actually generated from a model in the class that we consider.

*Department of Economics, The Ohio State University and NBER, Department of Economics, The Ohio State University and Stanford University, and Department of Economics, The Ohio State University, respectively. This paper is a revised version of NBER Working Paper, no. 2762, November 1988. We have benefited from comments received from seminar participants at Boston University, British Columbia, Claremont, Michigan, Ohio State, the Federal Reserve Board, Stanford, U.C.S.D., and the NBER. We thank Andy Abel, David Backus, Ben Bernanke, John Campbell, Steve Cosslett, Paul Evans, Jim Hamilton, Alan Gregory, Hu McCulloch, Robert Shiller, Larry Summers, and two anonymous referees for useful comments on earlier drafts. Cecchetti acknowledges financial support from the Salomon Bros. Center for the Study of Financial Institutions at NYU and the National Science Foundation.

[1] Poterba and Summers find negative serial correlation for stock returns over long horizons using monthly and annual data. Interestingly, Andrew Lo and A. Craig MacKinlay (1988) find that stock returns are positively correlated, using weekly observations.

[2] Sanford Grossman and Robert Shiller (1981) make this same point in showing that the "excess" volatility implied by variance bounds tests can be partly explained by risk aversion in a consumption beta model. More recently, Fischer Black (1988) has discussed the relation between mean reversion and consumption smoothing.

Specifically, we begin with a variant of the Lucas (1978) model of an exchange economy in which the parametric representations for preferences and the stochastic process governing the exogenous forcing variable (i.e., the endowment stream) admit a closed form solution to the asset pricing problem. We assume that the period utility function belongs to the constant relative risk aversion family. For these preferences, the coefficient of relative risk aversion is also the inverse of the elasticity of intertemporal substitution in consumption so that it is not possible to separate agents' tolerance for risk from their desire to have smooth consumption. Since our focus is on the serial correlation in asset returns implied by a model where agents confront an intertemporal consumption/investment decision, we believe that it makes more sense to interpret the concavity of the utility function in terms of the consumption smoothing motive.

The theory provides little guidance as to which time-series (i.e., consumption, output, or dividends) should serve as the endowment and from which to calibrate the model. That is because in the Lucas model, equilibrium consumption equals output, which also equals dividends. Since none of these time-series seem to be more appropriate than the others a priori, we examine each of the three series separately. The stochastic process followed by the growth rate of the endowment is assumed to follow James Hamilton's (1989) Markov switching model. This characterization of the forcing process has two important attributes. First, it allows us to model both the negative skewness and the excess kurtosis that is present in the growth rates of the raw data we employ.[3] And second, the Markov switching model admits a closed form solution to the intertemporal asset pricing problem. Neither of these objectives can be accomplished with standard linear ARIMA models.

The parameters of the Markov switching model are estimated by maximum likelihood employing annual observations on each of the series. Using these estimates, the empirical soundness of the Markov switching model is demonstrated by showing that it matches all three time-series well in the dimension of the mean, variance, skewness, kurtosis, and first-order serial correlation. Furthermore, comparisons of k-step ahead in-sample forecast errors of the Markov switching model with autoregressions reveal roughly similar predictive capabilities.

We then study the measures of mean reversion that have appeared in the literature. These are the variance ratio statistics used by Poterba and Summers and the long-horizon regression coefficients calculated by Fama and French.[4] First we calculate these statistics from historical data on returns drawn from the Standard and Poors' index. The asset pricing model is calibrated by setting the parameters of the endowment process equal to the maximum likelihood estimates. Using the calibrated equilibrium model, we construct Monte Carlo distributions for these statistics. Inferences regarding the model can then be drawn using classical hypothesis testing procedures and the Monte Carlo distributions as the null. We are primarily interested in two hypotheses. The first is the random walk model of stock prices, which is an implication of the Lucas model when agents have linear utility. The second hypothesis is that observed asset prices are determined in equilibrium but agents attempt to smooth their consumption. In this setting, asset returns can be negatively serially correlated even though they rationally reflect market fundamentals.

To summarize our results at the outset, we find that for all return horizons both the variance ratio statistics and the long-horizon regression coefficients calculated from the actual Standard and Poors' returns lie near

[3] The negative skewness of growth rates for many macroeconomic time-series, and hence their asymmetric behavior over the cycle, is also discussed in Salih Neftci (1984).

[4] We might also have examined variance bounds tests. But as John Campbell and Shiller (1988) point out, there is an equivalence between variance ratio tests of the type in Poterba and Summers and variance bounds tests pioneered by Shiller (1981).

the 60 percent confidence band of the median of the Monte Carlo distribution generated under the linear utility (random walk) model. When investors display only a moderate desire to smooth their consumption, these same statistics calculated from the data lie at or near the median of the Monte Carlo distribution. When we test the null hypothesis against a diffuse alternative, we cannot reject the random walk model at the standard 5 percent significance level, but the p-values of the test are much higher when the null distribution is generated assuming the utility function is concave. We conclude that much of the serial correlation in historical stock returns can be attributed to small sample bias. However, the serial correlation of returns found in the data better resembles that of the model when the utility function is concave.[5]

The remainder of the paper consists of three sections. Section I presents the solution to the equilibrium asset pricing problem of the Lucas model when agents have constant relative risk aversion preferences, and the endowment process is assumed to follow the Markov switching model. We include in this section the maximum likelihood estimates of the stochastic model, along with an evaluation of the model's performance. Section II describes the Monte Carlo experiments and reports the main results of the paper. The final section offers some conclusions.

I. The Equilibrium Model

A. *A Case of the Lucas Model*

Consider the economy studied by Lucas (1978) in which there are a large number of infinitely lived and identical agents and a fixed number of assets that exogenously produce units of the same nonstoreable consumption good. Let there be K agents and N productive units. Each asset has a single perfectly divisible claim outstanding on it, and these claims are traded in a competitive equity market. The first-order necessary conditions for a typical agents' optimization problem are

$$(1)\quad P_{j,t}U'(C_t) = \beta E_t U'(C_{t+1})\left[P_{j,t+1}+D_{j,t+1}\right],\quad j=1,2,\dots,N,$$

where P_j = the real price of asset j in terms of the consumption good.
$U'(C)$ = marginal utility of consumption, C, for a typical consumer/investor.
β = a subjective discount factor, $0<\beta<1$.
D_j = the payoff or dividend from the jth productive unit.
E_t = the mathematical expectation conditioned on information available at time t.

In equilibrium, per capita ownership of asset j is $1/K$. It follows that equilibrium per capita consumption, C, is the per capita claim to the total endowment in that period, $(1/K)\Sigma_{j=1}^{N}D_j$. Now, make this substitution in equation (1) and sum over j to obtain an equilibrium condition involving economy-wide or market prices and quantities on a per capita basis. That is,

$$(2)\quad P_tU'(D_t)=\beta E_tU'(D_{t+1})\left[P_{t+1}+D_{t+1}\right],$$

where $P_t \equiv (1/K)\Sigma P_{j,t}$ is the share of the market's value owned by a typical agent and $D_t=(1/K)\Sigma D_{j,t}$. Since each productive unit has only a single share outstanding and the number of productive units are fixed, these are the theoretical value-weighted market indices adjusted for population.

Let preferences be given by constant relative risk aversion utility:

$$U(C)=(1+\gamma)^{-1}C^{(1+\gamma)},$$

where $-\infty<\gamma\leq 0$ is the coefficient of relative risk aversion. Now (2) simplifies to a

[5] Myung Kim, Charles Nelson, and Richard Startz (1988) and Matthew Richardson (1988) have recently examined the issue of small sample bias in these tests of serial correlation in stock returns.

stochastic difference equation that is linear in PD^{γ}. That is,

$$(3)\quad P_tD_t^{\gamma} = \beta E_tP_{t+1}D_{t+1}^{\gamma} + \beta E_tD_{t+1}^{(1+\gamma)}.$$

Iterating (3) forward, the current market value, P_t, can be expressed as a nonlinear function of current and expected future payoffs,

$$(4)\qquad P_t = D_t^{-\gamma}\sum_{k=1}^{\infty}\beta^kE_tD_{t+k}^{\{1+\gamma\}}.$$

To obtain a closed form solution, we must specify the stochastic process governing (D_t), and this is done in Section I, Part C. We will refer to the exogenous forcing variable as dividends in the next two subsections. We do this because it helps to clarify the exposition, not because we restrict our attention to dividends when assessing the performance of the model. In fact, we consider alternatives as well.

B. *Characteristics of the Data*

The theory provides little guidance regarding the appropriate empirical counterpart to the exogenous forcing variable D. Because equilibrium consumption equals output, which also equals dividends, there are three natural variables to serve this role. We consider all three candidate time-series in real, per capita terms:[6] dividends, consumption, and GNP.

To choose an appropriate time-series model for the endowment process, it is useful to know some of the details of the data. Table 1 reports various summary statistics computed for the growth rates of dividends, GNP, and consumption. The data sources are described in the appendix. The following observations emerge from the table. Relative to a normal distribution, growth rates of the raw data are negatively skewed and have excess kurtosis. The coefficient of skewness is a measure of asymmetry, while the coefficient of kurtosis in excess of 3 implies that the distribution of the data is "fat tailed." The negative skewness indicates that, relative to a normal distribution, the data contain too many large negative values or "crashes." As Reitz (1988) suggests in his examination of the equity premium puzzle, these crashes are potentially important for studying the dynamics of asset returns.

The negative skewness is statistically significant at the 5 percent level for consumption and at the 1 percent level for dividends and GNP. Consequently, conventional time-series models, such as linear ARIMA models with Gaussian innovations, will be inappropriate. That is, ARIMA models with normal error terms can never explain nonzero third moments in the distribution of the raw data.

Excess kurtosis is also statistically significant at the 5 percent level for consumption and at the 1 percent level for dividends and GNP. Thus standard linear models will not capture this important characteristic of the data either. While conditionally normal but heteroskedastic models such as Robert Engle's (1982) ARCH model can give rise to fat tails, they cannot model asymmetry.

Our objective is to find a model that captures these important features of the data and at the same time admits a solution to the intertemporal asset pricing problem set forth in Section 1, Part A. Hamilton's (1989) Markov switching model meets both of these criteria.

C. *Hamilton's Markov Switching Model*

Hamilton (1989) has suggested modeling the trends in nonstationary time-series as Markov processes, and has applied this ap-

[6] The standard procedure in the literature has been to calibrate the endowment process to consumption (for example, Rajnish Mehra and Edward Prescott, 1985; Thomas Reitz, 1988; Shmuel Kandel and Robert Stambaugh, 1988; and George Constantinides, 1988, to name a few). It turns out that our results are robust to the particular time-series to which the endowment process is calibrated, whether it be consumption, dividends, or GNP. Because the variability of consumption, dividends, and GNP are very different from each other, the choice of the time-series to which the model is calibrated will have different implications for other aspects of the model such as the implied size and volatility of returns and the risk premium. The aim of this paper is quite modest, however, in that it seeks only to examine conclusions that can be drawn from serial correlation in returns. We make no claims that our model can match every dimension of the data (see Kandel and Stambaugh, 1988, who undertake a more ambitious project).

TABLE 1—SUMMARY STATISTICS FOR GROWTH RATES IN SAMPLE

	Consumption	Dividends	GNP
Mean	0.0184	−0.0038	0.0183
Std. Dev.	0.0379	0.1359	0.0547
Skewness	−0.4097[a]	−0.5979[b]	−0.7574[b]
Coefficient	(0.247)	(0.227)	(0.225)
Kurtosis	3.8750[a]	5.3048[b]	7.6630[b]
Coefficient	(0.495)	(0.455)	(0.451)
Minimum	−0.1044	−0.4673	−0.2667
Maximum	0.0989	0.4056	0.1662
First Autocorrelation	−0.067	0.134	0.390[c]
	(0.101)	(0.093)	(0.092)

Source: See data appendix.

Notes: Standard errors for the skewness and kurtosis coefficients are reported in parentheses and are computed under the null hypothesis that growth rates of the data are distributed as i.i.d. normal. Significance tests are based on E. S. Pearson and H. O. Hartley (1976), Tables 34.B and 34.C.

[a] Significantly different from normal at the 5 percent level.

[b] Significantly different from normal at the 1 percent level.

[c] Greater than two standard errors from zero.

proach to the study of post-World War II real GNP. One of the attractive features of this approach is its ability to model the asymmetry and the leptokurtosis reported in Table 1. Let d_t denote the logarithm of the endowment, D_t. The Markov switching model can be written as

$$d_t = d_{t-1} + \varepsilon_t + \alpha_0 + \alpha_1 S_{t-1}, \tag{5}$$

where $\{\varepsilon_t\}$ is a sequence of independent and identically distributed normal variates with zero mean and variance σ^2, and $\{S_t\}$ is a sequence of Markov random variables that take on values of 0 or 1 with transition probabilities,

$$\begin{aligned} (6) \qquad & \Pr[S_t = 1 | S_{t-1} = 1] = p, \\ & \Pr[S_t = 0 | S_{t-1} = 1] = 1 - p, \\ & \Pr[S_t = 0 | S_{t-1} = 0] = q, \\ \text{and} \qquad & \Pr[S_t = 1 | S_{t-1} = 0] = 1 - q. \end{aligned}$$

The endowment process thus follows a random walk in logarithms ($d_t = d_{t-1} + \varepsilon_t$) with stochastic drift ($\alpha_0 + \alpha_1 S_{t-1}$). As a normalization we restrict α_1 to be negative. The economy is said to be in a high-growth state or boom when $S = 0$, and in a low-growth state or depression when $S = 1$. The probability of a boom next period given that the economy currently enjoys a boom is q, while the probability of a depression next period given a current depression state is p. The probabilities of transition from boom to depression and depression to boom are then 1-q and 1-p, respectively. The endowment grows at the rate α_0 during a boom, and $\alpha_0 + \alpha_1$ during a depression. The process $\{S_t\}$ can be represented as a first-order autoregression with an autocorrelation coefficient of $(p + q - 1)$ that can be interpreted as a measure of persistence in the forcing process.

It is also useful to think of the process loosely within the following context. The theory relates dividends to asset prices. In actual economies, future nominal dividend payments are announced in advance so a good deal of next period's dividend growth is currently known. This is captured by the timing of the state in the Markov trend, and in the next subsection agents in the artificial economy will be assumed to observe the current state of the economy. From (5), the forecastable part of dividend growth during period $t-1$ is $\alpha_0 + \alpha_1 S_{t-1}$, which is revealed at $t-1$. The unforecastable part of real divi-

TABLE 2—MAXIMUM LIKELIHOOD ESTIMATES OF THE MARKOV TREND MODEL

$$y_{t+1} = y_t + \alpha_0 + \alpha_1 s_t + \varepsilon_{t+1}$$
$$\text{Prob}[s_{t+1} = 1 | s_t = 1] = p, \text{Prob}[s_{t+1} = 0 | s_t = 1] = 1 - p,$$
$$\text{Prob}[s_{t+1} = 0 | s_t = 0] = q, \text{Prob}[s_{t+1} = 1 | s_t = 0] = 1 - q,$$
$$\varepsilon \text{ i.i.d.} \sim N(0, \sigma^2).$$

Parameter	Consumption	Dividends	GNP
p	0.5279	0.1748	0.5096
	(1.985)	(0.832)	(2.034)
q	0.9761	0.9508	0.9817
	(46.525)	(40.785)	(76.705)
σ	0.0320	0.1050	0.0433
	(12.297)	(13.682)	(14.932)
α_0	0.0228	0.0171	0.0246
	(6.467)	(1.579)	(5.950)
α_1	−0.0926	−0.3700	−0.1760
	(−4.894)	(−6.548)	(−7.116)
$\Pr(S_t = 1)$	0.0482	0.0563	0.0360

Note: Asymptotic t-ratios in parentheses.

dend growth, ε_t, might be thought of as a combination of unanticipated inflation and productivity shocks.

We note at this point that it is the data and not the discretion of the investigator that will choose the regime. That is, when we calculate the Monte Carlo distributions implied by the model, the parameters $(\alpha_0, \alpha_1, p, q, \sigma)$ of the forcing process will be set equal to maximum likelihood estimates obtained from the data.

D. *Maximum Likelihood Estimates of the Markov Switching Model*

This section reports maximum likelihood estimates of the Markov switching model described above for annual dividends, GNP, and consumption. The model is nonlinear in the sense that the current minimum mean square error predictor of future values is a nonlinear function of current and lagged observations. Even though the state, S, is unobservable to the econometrician, given the normality assumption on the ε's, the parameters of the process, $(p, q, \alpha_0, \alpha_1, \sigma)$ can be estimated by maximum likelihood. The interested reader is directed to Hamilton (1989) for details on estimation or Pok-sang Lam (1988), who generalizes the Hamilton model.

The estimation results and some summary statistics are reported in Table 2. For the most part, the parameters are accurately estimated. When the economy is in a boom this year, the estimated probability that it continues in a boom next year is q. This is estimated to be 0.95 for dividends, and 0.98 for GNP and consumption. The estimates of growth during a boom, α_0, are 0.017, 0.025, and 0.023 for dividends, GNP, and consumption, respectively. When in a boom, the estimated probability of a transition to a negative growth state next period, $1 - q$, is 0.05 for dividends, and 0.02 for GNP and consumption. While in a depression state, expected growth, $\alpha_0 + \alpha_1$, is -0.35 for dividends, -0.15 for GNP and -0.07 for consumption.

The table also reports the unconditional probability of observing a depression state, $\Pr(S_t = 1)$. These are 0.056 for dividends, 0.036 for GNP, and 0.048 for consumption. In other words, we expect real dividends, the most volatile of the three series, to crash by one-third in roughly 7 of the 116 years of the sample. While this may seem surprising, it is consistent with the historical experience. For example, the dividend model estimates imply that, if the economy is currently in the bad state $(S_t = 1)$, the 95 percent confidence interval for growth in per capita real dividends is $(-0.14, -0.56)$. The same confidence interval given that the economy is in the good state ($S_t = 0$) is $(0.23, -0.19)$. Consequently, if dividends fall by 20 percent or more, we can be fairly certain that $S_t = 1$. Of the 116 years in the sample, 8 meet this criterion.[7]

Once the economy finds itself in a depression, the probability that it will be in a depression the following year, p, is estimated to be 0.1748 for dividends, 0.5096 for GNP, and 0.5279 for consumption.[8]

E. *Evaluating the Markov Switching Model*

To assess the quality of the Markov switching model, we now compare it with some popular alternatives. The upper panel of Table 3 reports the results of two diagnostic tests. The first is a test of the Markov switching model against the simple nested null hypothesis that the data follow a geometric random walk with i.i.d. innovations. Because the Markov switching model is not identified under the null of the geometric random walk, the likelihood ratio statistic does not have the standard chi-squared distribution. Therefore, we have tabulated the distribution of the pseudo-likelihood ratio statistic in order to perform this test. Our Monte Carlo experiment involved 1000 replications where we first drew samples of 116 for growth rates of GNP and dividends, and 96 for growth rates of consumption under the null of a normal distribution with variance set to values computed from the data. Next, we fitted the Markov switching model to this artificial data, and finally, we computed a standard likelihood ratio statistic as twice the difference in the maximized log likelihood values of the null and alternative models. As reported in the table, the weakest case is for consumption, where we reject the random walk at the 0.8 percent level. For both GNP and dividends, we observed fewer than two cases in 1000 where the LR statistic in the Monte Carlo experiment exceed the value obtained in the sample. Given the results in Table 1, where the data clearly reject the hypothesis that growth rates were drawn from a normal distribution, it is perhaps not surprising that a formal statistical test rejects the random walk model for the log-levels.

The second test reported in Table 3 is for symmetry of the Markov transition matrix, which implies symmetry of the unconditional distribution of the growth rates. This test examines the maintained hypothesis that $p = q$ against the alternative that $p < q$.[9] The table reports statistics that are asymptotically standard normal under the null. We can reject the hypothesis of symmetry at the 5 percent level in all three cases.

The lower panel of Table 3 reports the distributional characteristics for the Markov switching process implied by the estimates in Table 2. We report the population values of the mean, standard deviation, coefficient of skewness, coefficient of kurtosis, and first-order autocorrelation computed from the point estimates of the Markov switching model. The values implied by the model generally lie within two standard deviations of the sample values reported in Table 1.[10] The lone exception is the coefficient of kurtosis for GNP. We conclude that the Markov switching model can produce both the degree of negative skewness and the amount of kurtosis that are found in the data.[11]

[7] Real per capita dividends fell by more than 30 percent during 4 years, from 20 percent to 30 percent during four years, and by 10 percent to 20 percent during nine years of the sample.

[8] We note that the likelihood function is fairly flat for variations in p, particularly in the estimation of the dividend process. This is not surprising given the asymmetric behavior of dividends over the business cycle. That is, downturns have generally been short lived, lasting between four and six quarters. This makes it difficult to obtain a good estimate of p using annual observations. Hamilton does not encounter this problem since he estimates his model using *quarterly* GNP data.

[9] This is a one-sided test of symmetry against the alternative of negative skewness.

[10] It is worth noting that distribution of growth rates implied by the Markov switching model is always leptokurtic. Consequently, it is a natural candidate for modeling data that exhibit fat tails.

[11] The results in Table 3 demonstrate that the two-state model we employ is capable of matching the first four central moments of the data and the first-order

TABLE 3—EVALUATING THE MARKOV SWITCHING MODEL

	Data Set		
	Consumption	Dividends	GNP
L.R. Statistic[a]	11.39	17.27	27.87
p-value	(0.008)	(0.001)	(0.000)
Symmetry test[b]	1.72	3.71	1.89
p-value	(0.04)	(0.00)	(0.03)
	Distributional Characteristics of the Process Implied by the Estimates		
	Consumption	Dividends	GNP
Mean	0.0184	−0.0038	0.0183
Std. Dev.	0.038	0.135	0.054
Skewness Coefficient	−0.617	−0.965	−1.096
Kurtosis Coefficient	4.22	5.03	6.030
First Autocorrelation	0.140	0.05	0.179

[a] Based on a Monte Carlo experiment with 1000 replications.
[b] This is the test that $p = q$. The alternative hypothesis is $p < q$.

As a final test of the Markov switching model, we have compared its ability to forecast growth rates in the three series with that of first- and second-order autoregressions. Table 4 reports the in-sample root mean square forecast error at horizons from one to ten years for an AR(1), and AR(2), and the Markov switching model.[12] The results show that for dividends, the root mean squared forecast error of the Markov model is one-half that of the other models at all horizons. For consumption and GNP, the predictive ability of the three models is roughly the same. The Markov model marginally outperforms the autoregressions for consumption but marginally underperforms an AR(2) for GNP.

The purpose of this section has been to demonstrate that the Markov switching model is a reasonable alternative to obvious linear models in standard use. In addition to its ability to capture certain prominent features of the data that linear models cannot, the added attractiveness of the Markov switching model for our purposes is its analytical tractability. We conclude that while there is no obvious and clear winner, a credible case can be made for the Markov switching model. While a better model of the data (especially for GNP) might incorporate both regime switching and AR components, we have been unsuccessful in our attempt to solve the asset pricing problem when the endowment follows such a process. Our choice of the simple Markov switching model thus embodies a tradeoff between a model that completely matches the data and one that is tractable.

F. *Equilibrium Asset Prices*

Assume that the process driving the endowment is given by the Markov switching model of equations (5) and (6). We now obtain a solution to the asset pricing problem stated in Section I, Part A, using the

autocorrelation. By contrast, linear ARIMA models can match higher-order serial correlation, but not the higher-order moments. A three-state Markov switching model would have ten parameters and would in principle be able to match the first through sixth autocorrelation of the data in addition to the first four central moments. We have explored the three-state model but have thus far been unable to obtain sensible MLE parameter estimates with the annual data series studied here.

[12] Since these models are not nested, there is no natural statistical procedure for model selection.

TABLE 4—ROOT MEAN SQUARE FORECAST ERROR COMPARISONS USING LEVELS[a]

A. Consumption					Horizon					
Model	1	2	3	4	5	6	7	8	9	10
AR1	0.136	0.205	0.254	0.287	0.313*	0.340*	0.366	0.393*	0.415*	0.428*
AR2	0.136	0.205	0.257	0.288	0.314	0.341	0.367	0.396	0.417	0.431
MARKOV	0.131*	0.196*	0.249*	0.286*	0.315	0.342	0.366*	0.396	0.419	0.432

B. GNP					Horizon					
Model	1	2	3	4	5	6	7	8	9	10
AR1	0.324	0.594	0.808	0.962	1.055	1.093	1.109	1.126	1.186	1.241
AR2	0.318*	0.578*	0.790*	0.936	1.023	1.059	1.079	1.095*	1.154*	1.202*
MARKOV	0.351	0.608	0.796	0.936*	1.020*	1.054*	1.074*	1.102	1.167	1.220

C. Dividends					Horizon					
Model	1	2	3	4	5	6	7	8	9	10
AR1	0.017	0.021	0.025	0.029	0.033	0.038	0.042	0.047	0.052	0.057
AR2	0.031	0.033	0.034	0.037	0.040	0.044	0.047	0.051	0.055	0.060
MARKOV	0.009*	0.012*	0.014*	0.017*	0.020*	0.023*	0.026*	0.029*	0.032*	0.035*

[a] A * indicates lowest root mean square forecast error. Some entries appear the same due to rounding.

method of undetermined coefficients.[13] Conjecture the following solution:

$$(7) \qquad P_t = \rho(S_t)D_t.$$

The problem is to verify that (7) solves (3) and to find the function $\rho(S_t)$. To do this, substitute (7) into (3) to obtain

$$(8) \quad \rho(S_t)D_t^{\gamma} = \beta E_t D_{t+1}^{\gamma+1}[\rho(S_{t+1})+1].$$

Next, write (5) in levels,

$$(9) \qquad D_{t+1} = D_t e^{(\alpha_0+\alpha_1 S_t+\varepsilon_{t+1})}.$$

Now substitute (9) into (8) and note that ε is i.i.d. normal with variance σ^2 to obtain

$$(10) \quad \rho(S_t) = \beta e^{[\alpha_0(1+\gamma)+(1+\gamma)^2\sigma^2/2]} \times e^{[\alpha_1(1+\gamma)S_t]}E_t[\rho(S_{t+1})+1].$$

Because S_t can take on only two values, 0 or 1, (10) is a system of two linear equations in $\rho(0)$ and $\rho(1)$. The solution is given by

$$(11) \quad \rho(0) = \tilde{\beta}[1-\tilde{\beta}\tilde{\alpha}_1(p+q-1)]/\Delta,$$

$$(12) \quad \rho(1) = \tilde{\beta}\tilde{\alpha}_1[1-\tilde{\beta}(p+q-1)]/\Delta,$$

where $\tilde{\beta}=\beta e^{[\alpha_0(1+\gamma)+(1+\gamma)^2\sigma^2/2]}$, $\tilde{\alpha}_1=e^{\alpha_1(1+\gamma)}$, and $\Delta = 1-\tilde{\beta}(p\tilde{\alpha}_1+q)+\tilde{\beta}^2\tilde{\alpha}_1(p+q-1)$. This establishes that (7) is the solution to (3).[14]

A number of interesting features of the equilibrium price function emerge. First, asset prices are proportional to the endowment.[15] Second, the factor of proportionality depends on the inverse of the elasticity of intertemporal substitution and whether the

[13] Using formulas derived by Hamilton (1989), one can also evaluate the series given in equation (4) directly. The appendix to the working paper version of this paper shows that both methods yield the same answer.

[14] It is possible to show that as long as both $\rho(0) > 0$ and $\rho(1) > 0$, the transversality condition is met and the power series (4) converges. In addition, this solution technique can easily be generalized to the case of n-states in the mean and the variance.

[15] In the simple model studied here this implies that the price dividend ratio takes on one of two values, $\rho(0)$ or $\rho(1)$. This is a consequence of assuming that agents observe S_t. In the more realistic case in which S_t is unobserved and must be estimated, the price dividend ratio would be a continuous variable fluctuating between the two bounds of $\rho(0)$ and $\rho(1)$.

economy is currently in the high-growth state or low-growth state according to

$$\rho(0) \gtreqless \rho(1) \text{ as } \gamma \gtreqless -1.$$

The interpretation of this is straightforward. For a given level of the current endowment, suppose that the economy is known to be in a high-growth state ($S_t = 0$). By (6), this implies that the economy is likely to remain in a high-growth state into the future, and agents anticipate high future levels of the endowment. This has two effects on asset prices that work in opposite directions. First, there is an intertemporal relative price effect in which the higher expected future endowment implies a lower relative price of future goods. This induces agents to want to increase saving and to increase their demand for assets. The increased asset demand arising from this intertemporal relative price effect works to raise current asset prices. Working in the opposite direction is a substitution effect arising from agents' attempts to smooth their consumption. When the expected future endowment is high, the consumption smoothing motive leads agents to want to increase current consumption in anticipation of higher future investment income. To finance higher current consumption, they attempt to sell off part of their asset holdings, which in equilibrium results in falling asset prices.

Log utility ($\gamma = -1$) is a borderline case in which the intertemporal relative price effect and the consumption smoothing effect exactly cancel out. This can be seen, perhaps, more clearly from (4), in which the solution for $\gamma = -1$ is $P_t = (\beta/[1-\beta])D_t$. In this case, the factor of proportionality relating prices to dividends is a constant. When the utility function is less concave than it is in the log case, the intertemporal relative price effect assumes greater importance, so that $\rho(0) > \rho(1)$. In the limiting case of linear utility ($\gamma = 0$), the intertemporal relative price effect is all that matters since agents have no desire to smooth consumption. Conversely, when the utility function is more concave than is implied by log utility, the intertemporal consumption smoothing effect dominates the intertemporal relative price effect causing $\rho(1) > \rho(0)$.

From (5) and (7), equilibrium gross returns are computed as

$$(13) \quad \begin{aligned} R_t &= (P_t + D_t)/P_{t-1} \\ &= \{[\rho(S_t)+1]/\rho(S_{t-1})\} \\ &\quad \times \exp[\alpha_0 + \alpha_1 S_{t-1} + \varepsilon_t]. \end{aligned}$$

Notice that because the gross return depends on ε_t, it is a continuous random variable on $[0, \infty)$ and not a two-point process.

II. The Serial Correlation of Equilibrium and Historical Returns

In this section, returns obtained from the equilibrium model of Section I are used to generate Monte Carlo distributions of the variance ratio statistics used by Poterba and Summers and the regression coefficients calculated by Fama and French. These distributions are generated both for the case of linear utility and for a case in which the utility function is concave. They are then used to draw inference about the equilibrium model and the model driving the exogenous forcing variable. For a given value of the elasticity of intertemporal substitution, the model is calibrated to the estimated dividend, consumption, and GNP processes reported in Table 1. That is, the parameters of the forcing process, $(p, q, \sigma, \alpha_0, \alpha_1)$ are set to the values in the columns of Table 2, and each case is considered in turn. The subjective discount factor β is assumed to be 0.97 throughout.

The procedure is as follows: First, given p and q, we generate a sequence of 116 S_t's according to (6). Second, given σ, 116 independent draws from a normal distribution with zero mean and variance σ^2 are taken to form a sequence of ε_t's. Third, given α_0, α_1, β, γ, $\{s_t\}$, and $\{\varepsilon_t\}$, we generate a sample of 116 returns according to equation (13). For each sample of returns, the variance ratio and regression coefficients are calculated for horizons 1 through 10. This experiment is repeated 10,000 times. The tabulation of these calculations is the Monte Carlo distribution of the statistic from which we draw inference. The sample size of 116 is chosen to correspond to the 116 annual observations available in the actual Standard and Poors' returns. We calculate the median, 60 percent confidence intervals about the median of the distribution, and the p-value for the statistic under investigation. We refer to these as the "small" sample results. A median is also calculated from 10,000 time-series samples of 1160 returns each, to get an idea of the rate of convergence of the variance ratio or regression coefficient statistic to its true population value. We refer to this as the "large" sample median. The results we obtain when the model is calibrated to consumption is representative of the results for dividend and GNP models. To facilitate the exposition, the results for the consumption model are also displayed in figures. Each figure displays the small sample medians, the 60 percent confidence intervals about the median, and the point estimates calculated from the historical Standard and Poors' returns.

From the Monte Carlo distributions of the variance ratio statistic and the autocorrelation coefficient on returns, we can determine the likelihood that the estimates obtained from the historical data were drawn from the Monte Carlo distribution implied by equilibrium returns.

A. *Variance Ratios*

Let R_t be the one period real rate of return, and R_t^k be the simple k-period return. That is, $R_t^k = \Sigma_{j=0}^{k-1} R_{t-j}$. Poterba and Summers define the variance ratio for returns at the kth horizon as

$$VR(k) = \frac{\operatorname{Var}(R_t^k)}{k \operatorname{Var}(R_t)}. \tag{14}$$

It is easy to show that the variance ratio can be expressed in terms of the return's autocorrelations. That is,

$$VR(k) = 1 + \frac{2}{k} \sum_{j=1}^{k-1} (k-j)\rho_j, \tag{15}$$

where ρ_j is the jth autocorrelation of annual returns. When returns are serially uncorrelated, the variance ratio is equal to one for all k in large samples.[16] This is usually taken as the null hypothesis in tests of "market efficiency," corresponds to the case where stock prices follow a random walk, and is true in the equilibrium model of Section I only when investors have linear utility. Stock prices are said to be "mean reverting" if returns are negatively serially correlated and evidence of mean reversion is inferred from variance ratios that lie below unity. This is the finding of Poterba and Summers.

We consider the case of linear utility first. Figure 1 displays the results under linear utility for the model calibrated to the consumption process. Since these returns are uncorrelated by construction, all of the deviation of the median of the variance ratio's distribution from unity is due to small sample bias.[17] In the large sample ($T = 1160$), most of the bias has disappeared. It is also seen that the variance ratios calculated from the Standard and Poors' data fall within the 60 percent confidence interval of the Monte Carlo distribution.[18] The serial correlation of returns, and hence their predictability, is only apparent.

This result can be viewed in the same light as the business cycle in which recessions occur with random periodicity. Although real GNP may appear to be mean reverting, this does not imply that business cycle turning points are predictable. In the equilibrium model of asset prices, the exogenous forcing

[16] In small samples, as Poterba and Summers point out, the sample autocorrelations of returns are biased so $E[VR(k)] < 1$ even when returns are independent.

[17] It bears mentioning that even in the case of linear utility and the geometric random walk, our empirical distributions differ from those reported in Poterba and Summers and Fama and French. The reason is that we assume a probability model for the endowment process and study the dynamics of the returns implied by an equilibrium model, whereas these authors assume a probability model for the returns. Since the return is a nonlinear function of the endowment in our setup, these two approaches need not yield the same small sample results.

[18] These estimates of the variance ratios are smaller than those reported by Poterba and Summers because they make a bias correction assuming a null hypothesis of a homoscedastic random walk for asset prices. The bias correction is irrelevant for our purposes.

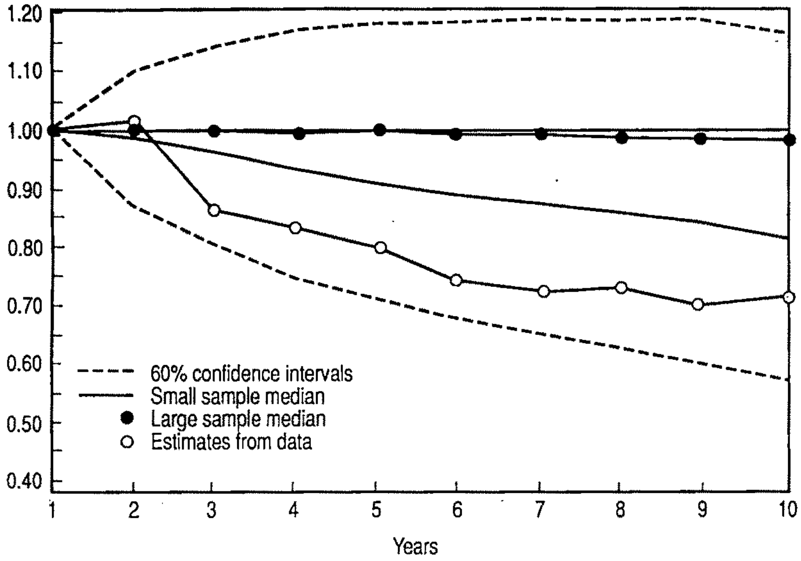


FIGURE 1. VARIANCE RATIOS. EQUILIBRIUM RETURNS GENERATED BY LINEAR UTILITY USING CONSUMPTION

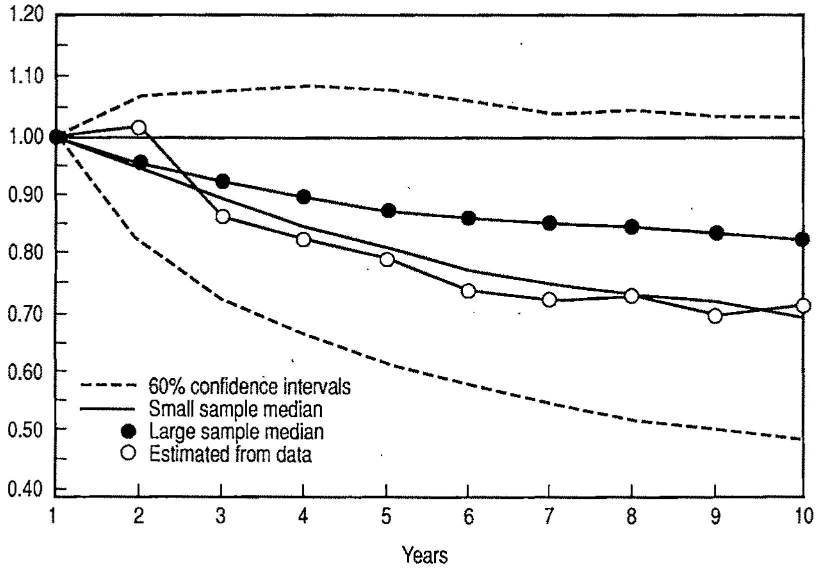


FIGURE 2. VARIANCE RATIOS. EQUILIBRIUM RETURNS GENERATED BY CONCAVE UTILITY USING CONSUMPTION

variable has a business cycle interpretation. The stochastic process of the forcing variable implies that a boom will eventually be followed by a recession, which will quickly be followed by a boom. Since equilibrium asset prices are proportional to the forcing variable, the appearance of mean reversion of asset prices is produced, but this does not mean that returns are predictable.

When agents' utility function is concave, the results are even more favorable to the model. Figure 2 reports the results of the

TABLE 5—VARIANCE RATIOS FOR HISTORICAL AND EQUILIBRIUM RETURNS

		Linear Utility			Concave Utility			
		$T=116$		$T=1160$	$T=116$		$T=1160$	
k	Actual	Median	p-Value	Median	Median	p-Value	Median	Population
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			Endowment Calibrated to Consumption:					
2	1.0137	0.9865	0.5779	1.0004	0.9462	0.6811	0.9523	0.9524
3	0.8664	0.9669	0.3049	1.0005	0.8941	0.4482	0.9212	0.9206
4	0.8351	0.9385	0.3301	0.9977	0.8511	0.4742	0.9006	0.8987
5	0.7978	0.9115	0.3227	0.9989	0.8181	0.4671	0.8812	0.8831
6	0.7459	0.8926	0.2903	0.9950	0.7782	0.4515	0.8707	0.8716
7	0.7259	0.8811	0.2877	0.9953	0.7576	0.4534	0.8613	0.8630
8	0.7363	0.8678	0.3307	0.9912	0.7380	0.4965	0.8528	0.8564
9	0.7102	0.8527	0.3238	0.9876	0.7264	0.4796	0.8474	0.8511
10	0.7242	0.8268	0.3737	0.9869	0.7083	0.5216	0.8391	0.8469
			Endowment Calibrated to Dividends:					
2	1.0137	0.9835	0.6151	0.9959	0.8721	0.8986	0.8786	0.8828
3	0.8664	0.9683	0.2374	0.9934	0.8166	0.6350	0.8266	0.8340
4	0.8351	0.9541	0.2420	0.9912	0.7746	0.6425	0.8021	0.8086
5	0.7978	0.9367	0.2401	0.9906	0.7478	0.6069	0.7848	0.7933
6	0.7459	0.9230	0.2042	0.9891	0.7286	0.5355	0.7742	0.7831
7	0.7259	0.9058	0.2168	0.9876	0.7103	0.5277	0.7654	0.7758
8	0.7363	0.8906	0.2582	0.9875	0.6957	0.5759	0.7595	0.7704
9	0.7102	0.8851	0.2490	0.9838	0.6821	0.5492	0.7547	0.7661
10	0.7242	0.8682	0.3000	0.9840	0.6718	0.5858	0.7496	0.7627
			Endowment Calibrated to GNP:					
2	1.0137	0.9782	0.5877	0.9966	0.9401	0.6632	0.9463	0.9490
3	0.8664	0.9403	0.3722	0.9936	0.8760	0.4868	0.9102	0.9153
4	0.8351	0.9075	0.3837	0.9911	0.8296	0.5074	0.8839	0.8923
5	0.7978	0.8806	0.3757	0.9859	0.7858	0.5168	0.8658	0.8761
6	0.7459	0.8590	0.3451	0.9850	0.7516	0.4911	0.8492	0.8643
7	0.7259	0.8385	0.3541	0.9778	0.7424	0.4800	0.8390	0.8554
8	0.7363	0.8252	0.3909	0.9773	0.7136	0.5309	0.8285	0.8486
9	0.7102	0.8120	0.3763	0.9770	0.6933	0.5229	0.8204	0.8432
10	0.7242	0.7857	0.4227	0.9730	0.6833	0.5487	0.8172	0.8389

Notes: Under linear utility, $\gamma = 0$. Under concave utility, γ is set to -1.7 for the consumption model, -1.4 for the dividend model, and -1.6 for the GNP model. $\beta = 0.97$ throughout. Column 1: Horizon, in years. Column 2: Variance ratios of historical Standard and Poors' returns. Column 3: Median of Monte Carlo distribution of variance ratios for 116 equilibrium returns generated with linear utility. Column 4: Percentage of Monte Carlo distribution having values less than the value in column 3. Column 5: Median of Monte Carlo distribution for variance ratios of 1160 equilibrium returns generated with linear utility. Column 6: Median of Monte Carlo distribution of variance ratios of 116 equilibrium returns generated with concave utility. Column 7: Percentage of Monte Carlo distribution with values less than the value in column 6. Column 8: Median of Monte Carlo distribution for variance ratios of 1160 equilibrium returns generated with concave utility. Column 9: Population variance ratio of equilibrium returns.

above calculations assuming concave utility with $\gamma = -1.7$ and the forcing process matched to consumption. Now the median of both the small and large sample distributions of the variance ratio statistics are well below 1.0 at every horizon. The median of the small sample ($T = 116$) distribution closely matches the variance ratios calculated from the annual returns on the Standard and Poors'.[19]

Table 5 reports results for the model calibrated to consumption, dividends, and GNP,

[19] When $\gamma = -2$, the model yields much more mean reversion than is in the data. The entire 60 percent confidence band lies below the sample values.

from which we make the following observations. First, the results for variance ratios are fairly robust to the choice of the time-series to which the model is calibrated. *P*-values between 0.2 and 0.8 imply that the sample variance ratios lie within the 60 percent confidence interval of the Monte Carlo distribution median. Thus, it can be seen that even under linear utility, the model cannot be rejected at conventional significance levels. From column 4, the smallest *p*-value is obtained at the six-year horizon in the dividend model (p-value $= 0.204$). When the sample size is increased tenfold, ($T = 1160$), most of the small sample bias disappears. When the utility function is concave, the median of the distribution matches up well with the values implied by the data as the *p*-values in column 7 are generally in the neighborhood of 0.5. A sizable small sample bias remains present, but as in the linear utility case, most of this bias vanishes if the sample is made ten times longer.

B. *Regression Coefficients on Returns of Varying Horizons*

Consider estimating the first-order serial correlation coefficient on τ-year returns by running the following regression:

$$R_{t,t+\tau} = a_\tau + b_\tau R_{t-\tau,t} + u_{t,t+\tau}, \qquad \tau = 1,2,\ldots,10 \text{ years}, \tag{16}$$

where $R_{t,t+\tau}$ is the continuously compounded real stock return from t to $t+\tau$. It is easy to show that the relation between the autocorrelations of one-period returns and the autocorrelation of the τ-period return is

$$b_\tau = \frac{\rho_1 + 2\rho_2 + \cdots + \tau\rho_\tau + (\tau-1)\rho_{\tau+1}}{\tau + 2(\tau-1)\rho_1 + 2(\tau-2)\rho_2} \times \frac{+ \cdots + 2\rho_{2\tau-2} + \rho_{2\tau-1}}{+ \cdots + 2\rho_{\tau-1}}.$$

Using monthly returns on the CRSP index, Fama and French find that the slope coefficient b_τ is negative for τ greater than one year. From this they infer that stock prices are mean reverting. We examine their result by computing the empirical distribution of these regression coefficients implied by the model in Section I.

We begin with the linear utility ($\gamma = 0$) case. Figure 3 displays results for the model calibrated to consumption, the small sample median, and 60 percent confidence intervals of the Monte Carlo distribution of the regression coefficient b_τ, the population values implied by the model, and the estimates obtained from the Standard and Poors' returns. Again, the deviation of the median of the small sample ($T = 116$) distribution from zero is due to small sample bias. This bias increases as τ gets larger, because the effective sample size, as measured by the number of independent pieces of information (nonoverlapping observations), decreases with τ. For example, at the ten-year horizon, there are only ten nonoverlapping observations available in the Standard and Poors' data, and six nonoverlapping observations available in the CRSP returns!

The median of the large sample distribution ($T = 1160$), on the other hand, is reasonably close to the true value of zero. The regression coefficients calculated from the Standard and Poors' data uniformly lie below the median of the small sample Monte Carlo distribution in the consumption model.

Figure 4 displays the details of the Monte Carlo distributions of the regression coefficients obtained from the equilibrium returns when $\gamma = -1.7$ in the consumption model. As in Figure 3, the regression coefficient uniformly lies within the 60 percent confidence interval of the median. The distance between the small sample medians and the actual estimates tend to be smaller here than when agents have linear utility.

Table 6 reports results for the model calibrated to consumption, dividends, and GNP, from which we make the following observations. As we found with the variance ratios, the results on the regression coefficients are robust to the series to which the model is calibrated. Under linear utility, the strongest evidence against the model comes when the model is calibrated to dividends at the two-year horizon (p-value $= 0.1099$). Most of the small sample bias vanishes when $T = 1160$.

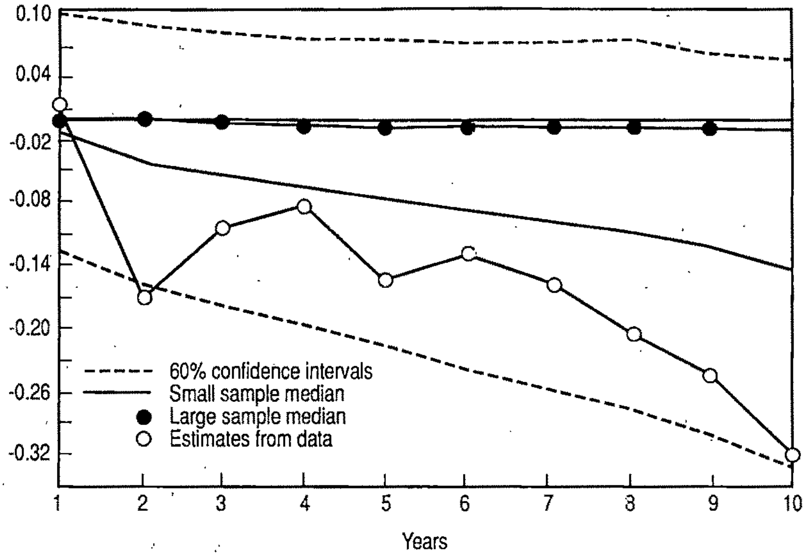


FIGURE 3. RETURN REGRESSION COEFFICIENTS. EQUILIBRIUM RETURNS GENERATED BY LINEAR UTILITY USING CONSUMPTION

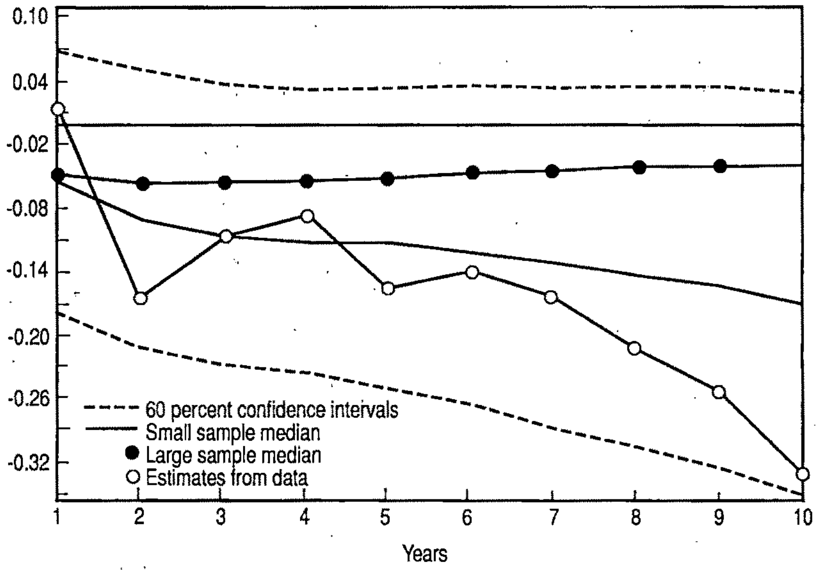


FIGURE 4. RETURN REGRESSION COEFFICIENTS. EQUILIBRIUM RETURNS GENERATED BY CONCAVE UTILITY USING CONSUMPTION

Table 6—Regression Coefficients for Historical and Equilibrium Returns

		Linear Utility			Concave Utility			
		$T=116$		$T=1160$	$T=116$		$T=1160$	
k	Actual	Median	p-Value	Median	Median	p-Value	Median	Population
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Endowment Calibrated to Consumption:								
1	0.0147	−0.0108	0.5747	0.0002	−0.0541	0.6788	−0.0466	−0.0475
2	−0.1671	−0.0416	0.1761	−0.0006	−0.0905	0.3039	−0.0555	−0.0564
3	−0.1059	−0.0543	0.3654	−0.0024	−0.1052	0.4989	−0.0532	−0.0531
4	−0.0836	−0.0658	0.4565	−0.0039	−0.1109	0.5650	−0.0510	−0.0470
5	−0.1530	−0.0758	0.3273	−0.0067	−0.1092	0.3940	−0.0474	−0.0409
6	−0.1345	−0.0866	0.3952	−0.0076	−0.1175	0.4629	−0.0430	−0.0357
7	−0.1601	−0.1000	0.3816	−0.0096	−0.1283	0.4335	−0.0405	−0.0314
8	−0.2098	−0.1108	0.3148	−0.0099	−0.1400	0.3586	−0.0365	−0.0279
9	−0.2488	−0.1267	0.2883	−0.0127	−0.1482	0.3119	−0.0362	−0.0251
10	−0.3246	−0.1472	0.2158	−0.0125	−0.1643	0.2277	−0.0330	−0.0227
Endowment Calibrated to Dividends:								
1	0.0147	−0.0129	0.6047	−0.0035	−0.1259	0.8992	−0.1215	−0.1171
2	−0.1671	−0.0254	0.1099	−0.0039	−0.1058	0.2943	−0.0865	−0.0840
3	−0.1059	−0.0383	0.3095	−0.0031	−0.0948	0.4681	−0.0645	−0.0609
4	−0.0836	−0.0472	0.4055	−0.0045	−0.0985	0.5379	−0.0511	−0.0473
5	−0.1530	−0.0587	0.2914	−0.0051	−0.1030	0.3827	−0.0434	−0.0386
6	−0.1345	−0.0756	0.3813	−0.0064	−0.1050	0.4322	−0.0393	−0.0325
7	−0.1601	−0.0919	0.3617	−0.0069	−0.1141	0.4052	−0.0356	−0.0282
8	−0.2098	−0.1078	0.3150	−0.0078	−0.1269	0.3451	−0.0342	−0.0248
9	−0.2488	−0.1275	0.2896	−0.0101	−0.1387	0.3124	−0.0311	−0.0222
10	−0.3246	−0.1410	0.2080	−0.0108	−0.1589	0.2321	−0.0289	−0.0200
Endowment Calibrated to GNP:								
1	0.0147	−0.0187	0.5862	−0.0022	−0.0602	0.6659	−0.0538	−0.0509
2	−0.1671	−0.0577	0.2209	−0.0070	−0.1094	0.3584	−0.0656	−0.0597
3	−0.1059	−0.0667	0.3988	−0.0084	−0.1181	0.5299	−0.0631	−0.0557
4	−0.0836	−0.0731	0.4734	−0.0117	−0.1170	0.5814	−0.0594	−0.0489
5	−0.1530	−0.0822	0.3337	−0.0121	−0.1188	0.4202	−0.0544	−0.0424
6	−0.1345	−0.0951	0.4132	−0.0126	−0.1227	0.4739	−0.0506	−0.0369
7	−0.1601	−0.1041	0.3840	−0.0128	−0.1322	0.4413	−0.0454	−0.0324
8	−0.2098	−0.1160	0.3213	−0.0133	−0.1409	0.3665	−0.0424	−0.0288
9	−0.2488	−0.1269	0.2819	−0.0154	−0.1458	0.3133	−0.0406	−0.0258
10	−0.3246	−0.1433	0.2082	−0.0136	−0.1657	0.2262	−0.0403	−0.0234

Notes: Under linear utility, $\gamma = 0$. Under concave utility, γ is set to -1.7 for the consumption model, -1.4 for the dividend model, and -1.6 for the GNP model. $\beta = 0.97$ throughout. Column 1: Horizon, in years. Column 2: Regression coefficients of historical Standard and Poors' returns. Column 3: Median of Monte Carlo distribution of regression coefficients for 116 equilibrium returns generated with linear utility. Column 4: Percentage of Monte Carlo distribution having values less than the value in column 3. Column 5: Median of Monte Carlo distribution for regression coefficients of 1160 equilibrium returns generated with linear utility. Column 6: Median of Monte Carlo distribution of regression coefficients of 116 equilibrium returns generated with concave utility. Column 7: Percentage of Monte Carlo distribution with values less than the value in column 6. Column 8: Median of Monte Carlo distribution for regression coefficients of 1160 equilibrium returns generated with concave utility. Column 9: Population regression coefficients of equilibrium returns.

The model matches the data more closely when the utility function is concave. From column 7 it can be seen that the regression coefficient for one-year returns for the dividend model lies near the 95 percent confidence bound. At the remaining horizons in the dividend model and at all horizons for the consumption and GNP models, the regression coefficients computed with the data uniformly lie within a 60 percent confidence interval of the median of the Monte Carlo distribution. The distance between the small

TABLE 7—THE SOURCE OF THE NEGATIVE SERIAL CORRELATION, MEDIAN OF DISTRIBUTION OF VARIANCE RATIOS OF RETURNS FOR MODEL CALIBRATED TO CONSUMPTION

		GRW	MSM	GRW	MSM	GRW	MSM	GRW	MSM
		$\gamma=0$	$\gamma=0$	$\gamma=0$	$\gamma=0$	$\gamma=-1.7$	$\gamma=-1.7$	$\gamma=-1.7$	$\gamma=-1.7$
k	Actual	$T=116$	$T=116$	$T=1160$	$T=1160$	$T=116$	$T=116$	$T=1160$	$T-1160$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
2	1.0137	0.9913 (0.50)	0.9865 (0.58)	0.9983 (0.70)	1.0004 (0.62)	0.9907 (0.60)	0.9462 (0.68)	0.9991 (0.69)	0.9523 (0.90)
3	0.8664	0.9732 (0.21)	0.9669 (0.30)	0.9953 (0.00)	1.0005 (0.02)	0.9758 (0.21)	0.8941 (0.45)	0.9964 (0.00)	0.9212 (0.22)
4	0.8351	0.9605 (0.22)	0.9385 (0.33)	0.9944 (0.00)	0.9977 (0.02)	0.9587 (0.22)	0.8511 (0.47)	0.9955 (0.00)	0.9006 (0.22)
5	0.7978	0.9458 (0.21)	0.9115 (0.32)	0.9937 (0.00)	0.9989 (0.01)	0.9474 (0.22)	0.8181 (0.47)	0.9927 (0.00)	0.8812 (0.18)
6	0.7459	0.9366 (0.17)	0.8926 (0.29)	0.9904 (0.00)	0.9950 (0.00)	0.9302 (0.18)	0.7782 (0.45)	0.9921 (0.00)	0.8707 (0.10)
7	0.7259	0.9208 (0.20)	0.8811 (0.29)	0.9922 (0.00)	0.9953 (0.00)	0.9227 (0.18)	0.7576 (0.45)	0.9912 (0.00)	0.8613 (0.09)
8	0.7363	0.9062 (0.25)	0.8678 (0.33)	0.9892 (0.00)	0.9912 (0.01)	0.9077 (0.18)	0.7380 (0.50)	0.9908 (0.00)	0.8528 (0.15)
9	0.7102	0.8887 (0.24)	0.8527 (0.32)	0.9884 (0.00)	0.9876 (0.01)	0.8938 (0.24)	0.7264 (0.48)	0.9874 (0.00)	0.8474 (0.12)
10	0.7242	0.8817 (0.29)	0.8268 (0.37)	0.9858 (0.00)	0.9869 (0.01)	0.8718 (0.29)	0.7083 (0.52)	0.9874 (0.00)	0.8391 (0.17)

Note: GRW is the geometric random walk model. MSM is the Markov switching model.

sample medians and the actual estimates tend to be smaller when agents have concave utility.

C. *Mean Reversion, Small Sample Bias, and Consumption Smoothing*

There are three features of our model that contribute to apparent mean reverting asset prices. They are (1) the nonlinearity of the endowment process (i.e., the presence of S_t), (2) pure small sample bias ($T=116$), and (3) consumption smoothing ($\gamma<0$). The role of consumption smoothing can be deduced from a comparison of columns 4 and 7 in Tables 5 and 6. For the values of γ that we work with, introducing concavity to the utility function shifts roughly an additional 15 percent to 20 percent of the variance ratio distribution and 1 percent to 10 percent of the regression coefficient distribution below the data. Table 7 reports additional results that allow us to separate these three influences and quantify the contribution of each in generating negatively serially correlated returns. While the table only includes evidence for the variance ratio statistic when the model is calibrated to consumption, these results are representative of the other cases that we do not report to economize on space.

We compare our results to those obtained when the endowment is assumed to follow a geometric random walk with drift and i.i.d. innovations with variance set equal to the sample values obtained from the consumption data. The table reports the median value of a Monte Carlo experiment with 10,000 replications, along with the percentage of the empirical distribution that lies below the value computed from the data.

To isolate the effects of the nonlinear endowment process, we first examine the implications for linear utility. A comparison of columns (3) and (4) reveals that the Markov switching model makes the small sample bias larger than it is in the case of the geometric random walk. While the variance ratio statistics are very near their population values of

TABLE 8—SERIAL CORRELATION IN EXCESS RETURNS, EQUILIBRIUM RETURNS IMPLIED BY THE MODEL CALIBRATED TO CONSUMPTION

		Linear Utility		Concave Utility		
		$T=116$		$T=116$		
k	Actual	Median	p-Value	Median	p-Value	Population
(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Variance Ratios:			
2	1.0492	0.9866	0.7048	0.9878	0.7423	0.9993
3	0.9300	0.9602	0.4712	0.9729	0.4366	0.9988
4	0.9121	0.9325	0.5080	0.9514	0.4678	0.9985
5	0.8639	0.9124	0.4721	0.9424	0.4184	0.9982
6	0.7848	0.8998	0.3937	0.9231	0.3308	0.9981
7	0:7246	0.8793	0.3543	0.9044	0.2846	0.9980
8	0.7123	0.8577	0.3736	0.8968	0.3021	0.9979
9	0.7038	0.8421	0.3924	0.8854	0.3245	0.9978
10	0.7148	0.8345	0.4363	0.8665	0.3842	0.9977
			Regression Coefficients:			
1	0.0511	−0.0133	0.6809	−0.0114	0.7209	−0.692e−03
2	−0.1193	−0.0407	0.2871	−0.0281	0.2229	−0.783e−03
3	−0.1021	−0.0558	0.3834	−0.0387	0.3330	−0.714e−03
4	−0.1230	−0.0663	0.3609	−0.0499	0.3224	−0.616e−03
5	−0.0973	−0.0778	0.4504	−0.0654	0.4268	−0.527e−03
6	0.0557	−0.0894	0.7838	−0.0816	0.7569	−0.454e−03
7	0.1094	−0.0952	0.8470	−0.0956	0.8401	−0.396e−03
8	0.0904	−0.1090	0.8298	−0.1033	0.8226	−0.349e−03
9	0.1202	−0.1248	0.8618	−0.1225	0.8527	−0.312e−03
10	0.0863	−0.1436	0.8331	−0.1342	0.8133	−0.281e−03

Notes: Column 1: Horizon, in years. Column 2: Statistics computed using historical excess returns. Column 3: Median of Monte Carlo distribution of statistics for 116 equilibrium excess returns generated with linear utility. Column 4: Percentage of Monte Carlo distribution having values less than the value in column 3. Column 5: Median of Monte Carlo distribution of statistics of 116 equilibrium excess returns generated with concave utility. Column 6: Percentage of Monte Carlo distribution with values less than the value in column 5. Column 7: Population statistics for equilibrium excess returns.

1.0 for $T=1160$, the distribution shifts farther down in a sample of $T=116$ when the endowment follows the Markov switching process. The results in column (5) show that for the Markov switching model, roughly 10 percent more of the empirical distribution lies below the data than for the geometric random walk model reported in column (1).

The effect of consumption smoothing alone can also be deduced from the table. Columns (7) and (9) of Table 7 replicate the well known result that if the endowment process follows a geometric random walk, concave utility cannot produce mean reversion. This can also be seen from our solution for returns in equation (13). When there is no switching of states, the function ρ is constant. Because the ε_t's are independent, returns are serially uncorrelated regardless of the elasticity of intertemporal substitution.

Columns (7) and (8) report on the importance of the interaction between consumption smoothing and the Markov switching model. Comparing column (8) to column (3) shows the importance of the combined impact of these two effects in small samples. By comparison with the model calibrated to the geometric random walk with linear utility, the nonlinear model with concave utility yields roughly an additional 25 percent of the empirical distribution below the data when $T=116$. Furthermore, the results in column (10) show that even in large samples, consumption smoothing implies negative serial correlation in returns when the endowment follows the Markov switching process.

D. *Excess Returns*

Table 8 reports results on mean reversion in excess returns when the model is calibrated to consumption.[20] Poterba and Summers also find evidence of mean reversion in excess returns. The top panel of the table shows the results for the variance ratio statistics. Column (2) of table reports the values computed from historical excess returns. These are only slightly larger than the values for real equity returns reported in Table 5. Although the model generates time-varying excess returns for the parameter values that we consider (i.e., the implied variance of excess returns are not zero), they do not exhibit very much serial correlation. However, the estimates in column (2) still lie well within the standard 95 percent confidence intervals of the median of the Monte Carlo distributions.

The bottom panel of the table reports the results for the regression coefficients although Fama and French do not report results for excess returns. From column (2), it can be seen that the coefficients computed from historical excess returns are substantially different from those computed with real equity returns, and are strikingly different from any of the results reported by Fama and French. Rather than becoming more and more negative as the horizon lengthens, they become positive at six years, and grow larger and larger. Nevertheless, because of small sample bias in the calculation of the regression coefficients at long horizons, the data still lie within a 90 percent confidence interval of the median of our Monte Carlo distributions.

III. Conclusion

This paper demonstrates that the findings of Poterba and Summers (1988) and Fama and French (1988), that stock prices are mean reverting, are consistent with an equilibrium model of asset price determination. The question we addressed was whether the empirically observed serial correlation properties of stock returns can be generated by an equilibrium model of asset pricing. Monte Carlo distributions of Poterba and Summers' variance ratio statistics and Fama and French's long-horizon return regression coefficients are generated using equilibrium returns derived from the Lucas (1978) model and Hamilton's (1989) Markov switching process governing consumption, dividends, and GNP. When economic agents care about smoothing their consumption, the equilibrium model implies that stock prices are mean reverting. It is possible that this is what was detected by Poterba and Summers and Fama and French. However, even with a linear utility function, the variance ratios and regression coefficients calculated with the historical Standard and Poors' returns data are not significantly different from the median of our Monte Carlo distribution. This latter result underscores the problem that 116 annual observations do not contain very much information when computing statistics based on returns at five- or ten-year horizons. Both the bias and the size of confidence intervals generated by sampling variation grow as the *effective* sample size gets smaller. The implication for testing the null against local alternatives is complementary to Summers' (1986) point that most tests of market efficiency have virtually no power against what he calls fad alternatives. Since we have shown that a properly constructed equilibrium model can generate rational assets prices that exhibit a good deal of negative serial correlation, it follows that, given the available data, the test of any fad model will have very little power against the rather wide class of equilibrium alternatives. More precise estimates and more powerful tests can only come through the passage of time and not by sampling the data more frequently.[21] If there had been a well-functioning asset market since the time of the Norman invasion (A.D. 1066) and we had all the necessary price and dividend data, then we might begin to distinguish among some of the competing theories. We conclude that

[20] The results with the model calibrated to dividends and GNP are similar and are not reported to save space.

[21] That is, in computing the autocorrelation of ten-year returns, what is needed is more ten-year time periods and not weekly or daily observations. All we can do is wait.

the evidence drawn from variance ratios and return regression coefficients are not sufficient to rule out equilibrium models. It is important to emphasize, however, that these results do not prove that the equilibrium model is true, since it is impossible to prove rationality or irrationality.

DATA APPENDIX

The dividend data are annual observations from the Standard and Poors' index from 1871 to 1985 deflated by the CPI. This is the Standard and Poors' historical data used by Poterba and Summers. Observations on returns and the CPI from 1871 to 1926 are from Jack Wilson and Charles Jones (1987) and Wilson and Jones (1988), respectively. From 1926 to 1985, the data are from Roger Ibbotson and Rex Sinquefeld (1988). Observations on nominal dividends are those used by Campbell and Shiller (1987). We use these data as a benchmark because they represent the longest available time-series, and we believe that the characteristics of these data are representative of equity returns and dividend disbursements in general. Also, the Standard and Poors' index is one of the data sets used by Poterba and Summers, so a direct comparison can be made with some of their results. We follow both Poterba and Summers and Fama and French in deflating by the CPI.

The risk-free rate series is the *ex post* real return to holding a one-year U.S. Treasury security, or the equivalent, computed by subtracting realized inflation from the nominal interest rate. The nominal interest rate series was computed using data from four separate sources. The constructed series is intended to come as close as possible to a measure of the yield to maturity on a one-year U.S. security. For the period from 1920 to 1929, the basic data are drawn from the column giving the compound annual return from rolling "3- to 6-month Treasury notes and certificates," Table No. 122, page 460, of the *Banking and Monetary Statistics of the United States*. For a given year, the one-year yield was computed by assuming that the security was rolled over in July. For the period from 1930 to 1950, the data are the one-year yield for December of the previous year reported in the appendix to Stephen Cecchetti (1988). From 1951 to 1985, the data are the one-year zero coupon yield reported in J. Huston McCulloch (1988). For the period from 1871 to 1919, there is no direct information on government yields. To obtain a consistent series, we began by regressing the one-year yields from 1920 to 1987 (as described above) on the commercial paper rate constructed by Campbell and Shiller (1987). Since the commercial paper rate series extends back to 1871, we were able to estimate the implied government yield as the fitted values from this regression.

The real GNP data are constructed by combining data from 1869 to 1928 from Christina Romer (1989) with data from 1909 to 1928 from Romer (1985), and observations from 1929 to 1985 from the National Income and Product Accounts.

The consumption data are constructed by splicing the Kendrick consumption series reported in Nathan Balke and Robert Gordon (1986), from 1889 to 1928, with the National Income and Product Accounts data from 1928 to 1985. This series is the longest available series on aggregate personal consumption expenditure we are aware of.

In order to express quantities in per capita terms, we divided each time-series by annual population estimates. The estimates used are as follows. From 1869 to 1938 the data are from the *Historical Statistics of the United States*, Series A7 from 1869 to 1928 (with the data in footnote 1 for 1917 to 1919), and Series A6 from 1929 to 1938. From 1938 to 1985 the data are from the *Economic Report of the President*, 1989, Table B-31.

REFERENCES

Balke, Nathan S. and Gordon, Robert J., "Appendix B. Historical Data," in R. J. Gordon, ed. *The American Business Cycle*, Chicago: University of Chicago Press for NBER, 1986.

Black, Fischer, "Mean Reversion and Consumption Smoothing," NBER Working Paper, no. 2946, April 1989.

Campbell, John Y. and Shiller, Robert J., "Cointegration and Tests of Present Value Models," *Journal of Political Economy*, October 1987, *95*, 1062–88.

______ and ______, "Stock Prices, Earnings and Expected Dividends," *Journal of Finance*, July 1988, *43*, 661–76.

Cecchetti, Stephen G., "The Case of the Negative Nominal Interest Rates: New Estimates of the Term Structure of Interest Rates During the Great Depression," *Journal of Political Economy*, December 1988, *96*, 1111–41.

Constantinides, George M., "Habit Formations: A Resolution of the Equity Premium Puzzle," mimeo., Department of Finance, Graduate School of Business, University of Chicago, October 1988.

Engle, Robert F., "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, July 1982, *50*, 987–1007.

Fama, Eugene F. and French, Kenneth R., "Permanent and Temporary Components of Stock Prices," *Journal of Political Economy*, April 1988, *96*, 246–73.

Grossman, Stanford J. and Shiller, Robert J., "The Determinants of the Variability of Stock Market Prices." *American Economic Review*, May 1981, *71*, 222–27.

Hamilton, James D., "A New Approach to the

Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, March 1989, *57*, 357–84.

Ibbotson, Roger G. and Sinquefield, Rex A., *Stocks, Bonds, Bills, and Inflation 1988 Yearbook*, Chicago: Ibbotson Associates, 1988.

Kandel, Shmuel and Stambaugh, Robert F., "Modeling Expected Stock Returns for Long and Short Horizons," mimeo., Department of Finance, Wharton School, University of Pennsylvania, December 1988.

Kim, Myung Jig, Nelson, Charles R. and Startz, Richard, "Mean Reversion in Stock Prices? A Reappraisal of the Empirical Evidence," NBER Working Paper, no. 2795, December 1988.

Lam, Pok-sang, "The Generalized Hamilton Model: Estimation and Comparison with Other Models of Economic Time-Series," mimeo., The Ohio State University, 1988.

Leroy, Stephen F., "Risk Aversion and the Martingale Property of Stock Prices," *International Economic Review*, June 1973, *14*, 436–46.

Lo, Andrew W. and MacKinlay, A. Craig, "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, Spring 1988, *1*, 41–66.

Lucas, Robert E. Jr., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, *66*, 1429–45.

McCulloch, J. Huston, "Appendix II: U.S. Term Structure Data, 1946–1987," NBER Working Paper, no. 2341, August 1987.

Mehra, Rajnish and Prescott, Edward C., "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, March 1985, *15*, 145–61.

Michener, Ronald W., "Variance Bounds in a Simple Model of Asset Pricing," *Journal of Political Economy*, February 1982, *90*, 166–75.

Neftci, Salih N., "Are Economic Time Series Asymmetric Over the Business Cycle?" *Journal of Political Economy*, April 1984, *92*, 307–28.

Pearson, E. S. and Hartley, H. O., *Biometric Tables for Statisticians*, London: Biometrika Trust, 1976.

Poterba, James M. and Summers, Lawrence H., "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics*, October 1988, *22*, 27–59.

Reitz, Thomas A., "The Equity Premium: A Solution," *Journal of Monetary Economics*, July 1988, *22*, 117–33.

Richardson, Matthew, "Temporary Components of Stock Prices: A Skeptic's View," mimeo., Graduate School of Business, Stanford University, April 1988.

Romer, Christina D., "World War I and the Post-War Depression: A Reinterpretation Based on Alternative Estimates of GNP," *Journal of Monetary Economics*, July 1988, *22*, 91–116.

______, "The Pre-War Business Cycle Reconsidered: New Estimates of Gross National Product, 1869–1918," *Journal of Political Economy*, February 1989, *97*, 1–37.

Shiller, Robert J., "Do Stock Prices Move Too Much to Be Justified by Subsequent Movements in Dividends?" *American Economic Review*, June 1981, *71*, 421–36.

Summers, Lawrence H., "Does the Stock Market Rationally Reflect Fundamental Values?" *Journal of Finance*, July 1986, *41*, 591–601.

Wilson, Jack W. and Jones, Charles, "A Comparison of Annual Common Stock Returns: 1871–1925 with 1926–1985," *Journal of Business*, April 1987, *60*, 239–58.

______ **and** ______, "Inflation Measure for the Period," mimeo., Department of Economics and Business, North Carolina State University, September 1988.

Board of Governors of the Federal Reserve System, *Banking and Monetary Statistics, 1914 to 1940*, Washington: 1943.

Council of Economic Advisers, *Economic Report of the President*, Washington: USGPO, 1989.

U.S. Department of Commerce, Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970, Bicentennial Edition*, Washington: USGPO, 1977.

______, **Bureau of Economic Analysis,** *The National Income and Product Accounts of the United States, 1929–82*, Washington: USGPO, 1986.

Is the European Community an Optimal Currency Area? Optimal Taxation Versus the Cost of Multiple Currencies

By MATTHEW B. CANZONERI AND CAROL ANN ROGERS*

We propose a view of optimal currency areas that is based on the principles of public finance. Inflation taxes are distortionary, and an optimal spreading of tax distortions may require high inflation in one region and low inflation in another. Each region would need its own currency to do this. On the other hand, multiple currencies imply valuation and currency conversion costs, which impede trade between regions. This tradeoff is explored in the context of the European Community's debate over a common currency, using a two-country variant of Lucas and Stokey's cash-in-advance model. (JEL 430)

The European Community seems to be moving rapidly toward economic integration. Barriers to trade in goods and financial assets are to come down by 1992. However, the fate of the European Monetary System seems to be in some doubt, especially in light of the elimination of capital controls. Without capital controls, the maintenance of fixed exchange rates may require more convergence in monetary policy and inflation rates than has yet been observed. The "Delors Report" has gone so far as to propose a common currency for the EC.[1] But is the EC an optimal currency area? We would like to suggest a new way of looking at this question.

Over the last fifteen years, arguments for fixed exchange rates have focused on stabilization problems created by sticky prices or on inflation problems caused by policymakers' lack of credibility.[2] Sticky prices were thought to imply adjustment costs, and the case for fixed rates was analyzed along the lines of William Poole (1970) or Canzoneri and Gray (1985). Alternatively, it was argued that "weak" policymakers could combat inflationary expectations by tying their currencies to a more stable currency, and the case for fixed rates was analyzed along the lines of Robert Barro and David Gordon (1983). In this paper, however, we eschew all problems associated with sticky prices or the policymakers' credibility. We do not mean to suggest that these problems are unimportant; we simply want to focus attention on a neglected aspect of the optimal currency area question.

Our purpose is to present a view of optimal currency areas that is based on the principles of public finance.[3] Seignorage is

*Department of Economics, Georgetown University, Washington, D.C. 20057. We wish to thank Hans Genberg, Kent Matthews, Bennett McCallum, Patrick Minford, Anne Sibert, Alan Stockman, Lars Svensson, Carlos Vegh, Paul Wood, and an anonymous referee for their helpful comments. We would also like to thank seminar participants at the Center for Economic Research (Tilburg), the Institute for International Economic Studies (Stockholm), the National Bureau of Economic Research, the University of Chicago, Erasmus University, Georgetown University, the University of Houston, Indiana University, and New York University. However, we alone are responsible for what follows.

[1] See Jacques Delors (1989). Jacques Delors is President of the European Commission. The Committee for the Study of Economic and Monetary Union prepared the Delors report for the European Council.

[2] Hans Genberg (1988) provides an excellent discussion of this literature and relates it to earlier work; however, he focuses almost exclusively on small countries. Matthew Canzoneri and Jo Anna Gray (1985) and Matthew Canzoneri and Dale Henderson (1988) extend the discussion to large countries, using a game theoretic approach.

[3] We are following a research agenda suggested by Robert Lucas (1986). Indeed, a macroeconomic/public finance literature on the international aspects of fiscal policy is already developing; see Patrick Kehoe (1987), Jacob Frenzel and Assaf Razin (1988), Carlos Vegh and

one tax among many that can be imposed to raise revenue. As Table 1 shows, southern European countries raise between 5 and 10 percent of their revenues through seignorage, while northern European countries hardly depend upon seignorage at all.[4] Why do tax structures differ across Europe? The public finance literature provides one answer. Tax rates should be set to spread the distortions that taxes create; the marginal disutility from the last revenues raised should be equalized across all revenue sources. Optimal tax rates will depend upon characteristics of the activities being taxed, including collection costs; goods and services that are easily taxed in one region may be very difficult to tax in another. There is no reason to think that the optimal inflation tax for Germany will be the same as that for Italy. For that matter, there is no reason for the inflation tax in southern Italy to be the same as that in northern Italy. In any case, the implications of the principles of public finance are clear. Regions that require the same inflation tax may form an optimal currency area.

TABLE 1—SEIGNORAGE AS A PERCENT OF TAX REVENUE, 1979–1988

Greece	9.1	Belgium	0.4
Italy	6.2	France	1.3
Portugal	11.9	Germany	0.8
Spain	5.9	U.K.	0.5

Note: This table was drawn from Drazen (1988), Table 1. Drazen's table was in turn drawn from tables in Giavazzi (1988).

Since the characteristics that determine optimal tax rates may vary widely across countries, or even across regions within a country, one might expect that the public finance view will call for many different inflation taxes and many small currency areas. As Robert Mundell (1961) has observed, this does not accord with a long held view that trade in many currencies is intrinsically costly, so that currency areas should be large. Mundell writes (p. 662):

> ...It will be recalled that the older economists of the nineteenth century were internationalists and generally favored a world currency. ...Mill, like Bagehot and others, was concerned with the costs of valuation and money-changing,..., and it is readily seen that these costs tend to increase with the number of currencies. Any given money qua numeraire or unit of account fulfills this function less adequately if the prices of foreign goods are expressed in terms of foreign currency and must then be translated into domestic currency prices. Similarly, money in its role of medium of exchange is less useful if there are many currencies; although the costs of currency conversion are always present, they loom exceptionally large under inconvertibility or flexible exchange rates.

Similar notions have been expressed by those who favor a continuation of the EMS or the creation of a single European currency, though the actual cost of multiple currencies has never been measured.[5]

Pablo Guidotti (1988) and Roberto Chang (1988). Gregory Mankiw (1987) and Vittorio Grilli (1989) report some empirical evidence that U.S. and European inflation rates have been used for intertemporal tax spreading; Grilli (1989) also finds some evidence for the intratemporal tax spreading discussed here, if allowance is made for constraints imposed by the EMS.

Rudiger Dornbusch (1988a,b) has already argued that "countries for whom the efficient tax structure implies the use of an inflation tax...should not merge with others for whom zero inflation is the policy objective." He suggests that even the small loss of seignorage revenue that has resulted from the convergence of European inflation rates can have explosive consequences for the level of debt when the real rate of interest exceeds the rate of growth.

[4] Inflation rates did fall in southern Europe during the EMS period: However, the southern European countries also increased reserve requirements during this period. This broadened the seignorage tax base and maintained revenues to some extent; see Allan Drazen (1988) and Francesco Giavazzi (1988).

It seems unlikely that the southern European countries will be able to maintain their higher reserve requirements once banking services are allowed across borders. The 1992 agenda may put more fiscal pressures on the southern European countries than we have seen to date. Again, the comments of Dornbusch in fn. 3 are relevant.

[5] Some have tried to measure the effects of exchange rate volatility on trade flows; see Martin Bailey and George Tavlas' (1988) recent summary of this literature. Few have found these effects significant. However, exchange rate volatility would not seem to be a good proxy for the concerns Mundell and others had in mind.

Thus, we come to the factors we want to consider in drawing the boundaries of an optimal currency area. Multiple currencies allow separate inflation rates in each country or region, and this flexibility may well produce tax spreading efficiencies. However, these benefits must be weighed against the valuation and currency conversion costs that come with a multiplicity of currencies.

To illustrate this tradeoff, we construct a two-country model of what we call the "EC." We give "Italy" tax spreading reasons to opt for high inflation; we model "Germany" as a low inflation country. However, we also model valuation and currency conversion costs, costs that can be eliminated if Italy and Germany adopt a common currency. The EC is an optimal currency area if the costs of multiple currencies outweigh the benefits of tax spreading.

In Section I, we outline our model of the EC. The model is essentially a two-country version of the model used by Robert Lucas (1987) and Robert Lucas and Nancy Stokey (1987). Each country is populated by infinitely lived households that consume home cash goods, home credit goods, and foreign credit goods; the cash good is subject to a cash in advance constraint. Each country has a given level of public spending that must be financed through seignorage or a value added tax. We model only one asymmetry between the countries: Italy has a black market that cannot be taxed directly. Germany has no black market, and all German production is subject to the VAT. The black market is modeled as a cash good. So, Italy can use inflation to tax its black economy indirectly.[6]

The model comes in two versions. In the first, each country has its own currency, and doing business with two currencies wastes resources; it costs more to produce a credit good that is sold to a foreigner. The extra costs reflect the time and effort it takes to quote prices in a foreign currency and to convert the proceeds back into domestic currency. In the second version of the model, Europe has a common currency, and selling to a foreigner entails no additional costs.

TABLE 2—THE IMPORTANCE OF THE BLACK ECONOMY

	Percent of GNP in Italy	Percent of GNP in Germany
Source:		
Kent Matthews	15 to 20	7 to 10
de Grazia (1984)	20 to 25	2
Contini (1982)	14 to 20	–
Langfeldt (1983)	–	5 to 10

Note: Matthew's estimates, based on his reading of a number of studies, were given to us in private conversation.

In Section II, we will give a formal definition of the optimal currency area problem. The EC policymaker's problem is to choose the number of currencies (and to set the available tax rates) so as to maximize a weighted sum of the utilities of Italian and German households.

In Section III, we find the best outcome that can be achieved if two currencies are retained. With two currencies, the EC policymaker has enough taxes to achieve an efficient outcome. The optimal tax rates have to make the relative price of cash and credit goods in each country equal to the rate of transformation in production. In Italy, the inflation tax on cash goods (the black economy) must balance the VAT on credit goods. Germany's VAT is not distortionary, since it falls on both cash goods and credit goods. However, the relative price on cash and credit goods will be distorted in Germany if the nominal interest rate is greater than zero, since money must be held instead of bonds to purchase a cash good. The optimal inflation rate in Germany is equal to minus the rate of time preference, as Friedman has suggested. From a public finance point of view, the EC is not an optimal currency area. Any EMS-like constraint that makes Italian and German inflation rates converge must decrease welfare in one country or the other.

The relevance of our modeling of the tax spreading problem depends on the importance of the black market in Italy. Table 2 presents estimates of the relative size of the black economy in various countries; the esti-

[6] This way of modeling black markets was suggested to us by Paul Wood in the context of Latin American countries. Robert Barro (1987) has also noted that the inflation tax may be the only way of taxing the black economy.

mates vary widely, but Italy's black economy appears to be two or three times the size of black economies in other European countries. One might conclude that Italy does indeed have more reason to resort to the inflation tax. Of course, we do not mean to suggest that the black economy is Italy's only reason for adopting a high inflation rate. Similar results would obtain if Italy simply faced higher tax collection costs on cash goods than Germany. And more generally, Italy could differ from Germany in any number of ways that would result in a higher optimal inflation rate.

In Section IV, we ask if the EC would be better off with a common currency. The costs associated with a multiplicity of currencies would be eliminated, but Italy and Germany could no longer run the independent inflation rates they need for efficient tax spreading. The answer one might expect is that it depends on the size of the valuation and currency conversion costs: if the costs are high, then the EC is an optimal currency area; if the costs are small, then separate currencies should be maintained for tax spreading. We show that this answer is correct as long as home goods are not close substitutes for imported goods. High valuation and currency conversion costs drive out international trade. If home goods are close substitutes for imports, then the switch to home goods involves no great loss in utility; in this case, it may pay to keep the tax spreading flexibility afforded by multiple currencies, even if the extra costs do drive out trade.

While these arguments are relatively straightforward, our model of the EC is rather complex. This complexity makes more detailed results hard to obtain. Therefore, in Section V, we present some numerical examples that suggest the following propositions: (1) Very small valuation and currency conversion costs can make the EC an optimal currency area. (2) High levels of government spending make the EC less likely to be an optimal currency area. (3) Openness makes the EC more likely to be an optimal currency area. While the general validity of these propositions has yet to be demonstrated, our examples are suggestive and the propositions are plausible.

Finally, in Section VI, we discuss the limitations of our analysis. We also make suggestions for future research.

I. A Simple Model of the EC

The model comes in two versions. We begin with the multiple currency version, and then we explain how it has to be modified to get the single currency version.

There are only two countries in our stylized model of the EC: Italy and Germany. Italy is inhabited by identical, infinitely lived households whose utility depends upon consumption of an Italian cash good, x, an Italian credit good, y, and a German credit good, z:

$$(1) \qquad U=\sum_{t=0}^{\infty}\left[\frac{1}{1+\delta}\right]^{t}u(x_t,y_t,z_t).$$

The function $u(x,y,z)$ is increasing, strictly concave and twice continuously differentiable; in addition, $\partial u(0,y,z)/\partial x=\partial u(x,0,z)/\partial y=\partial u(x,y,0)/\partial z=\infty$. Germany is similarly populated by identical, infinitely lived households whose utility depends on consumption of a German cash good, x^*, a German credit good, y^*, and an Italian credit good, z^*:

$$(1^*) \qquad U^*=\sum_{t=0}^{\infty}\left[\frac{1}{1+\delta}\right]^{t}u(x_t^*,y_t^*,z_t^*).$$

*'s denote German variables, but the utility functions in Italy and Germany are the same.

Households in each country are endowed with n units of labor each period. This labor can be converted one for one into home cash goods or home credit goods. Since Italy and Germany have separate currencies, producers incur extra resource costs when they sell to foreigners: $c>1$ units of labor are needed to produce one credit good sold to a foreigner. Thus, the resource constraints for Italy and Germany are

$$(2) \qquad x_t+y_t+cz_t^*+g=n$$

and $$x_t^*+y_t^*+cz_t+g=n.$$

Government spending, g, is exogenously determined and the same in Italy and Germany.

The cash in advance framework and the distinction between cash goods and credit goods are borrowed from Lucas.[7] Each period begins with a financial exchange and ends with a goods market. In the financial exchange, households pay for the credit goods they purchased last period and acquire the cash they will need in the goods market that follows; governments collect taxes, pay interest on debt, and issue new debt (in the form of money or bonds). In the goods market, households divide into worker-shopper pairs. Workers produce cash goods and credit goods; shoppers visit the workers of other households and buy goods. Shoppers must pay cash for the cash goods; credit goods are invoiced in the shopper's currency for payment in the next financial exchange.

Now we can be more explicit about the extra costs workers incur when there are two currencies. Credit goods are, by assumption, invoiced in the buyer's currency, and in the next period the payment for them comes in the buyer's currency.[8] A worker incurs two additional costs in selling credit goods to a foreigner. First, the worker has to value the good in another unit of account. Second, the foreign currency received must be converted into domestic currency, so that the household can purchase domestic cash goods. Both of these activities take time and effort, so we have modeled them as direct resource costs. They are our model's incarnation of the valuation and currency conversion costs described by Mundell and others.

Certain arbitrage conditions are immediate, and it will facilitate our discussion to take advantage of them. A worker will charge the same price for cash goods and home credit goods, since each takes one unit of labor to produce and the sale of either results in cash that cannot be used until the next financial period. Let p_t be the lira price of the Italian goods, and let p_t^* be the mark price of the German goods. It takes c units of labor to produce a credit good for a foreigner, and the worker wants sales to all customers to result in the same cash value in the next financial exchange; so, letting e_t be the lira price of marks,

$$(3)\quad p_t^z = e_{t+1}p_t^*c \text{ and } p_t^{z*} = (p_t/e_{t+1})c$$

are the prices charged to Italian and German shoppers for foreign credit goods.

An Italian household's budget constraint for the financial exchange is

$$(4)\quad m_t + b_{It} + e_t b_{Gt} + \theta_t p_t(y_t + q_t z_t) \le R_{t-1}b_{It-1} + e_t R_{t-1}^* b_{Gt-1} + p_{t-1}(n - y_{t-1} - q_{t-1}z_{t-1}).$$

On the left-hand side, we have the household's acquisition of money and bonds and its tax payments. "I" subscripts denote Italian government bonds, which pay the gross nominal return R, and "G" subscripts denote German government bonds, which pay R^*. The Italian government levies a value added tax, θ_t, on the household's purchases of domestic and foreign credit goods; q_t ($\equiv p_t^z/p_t$) is the relative price of the foreign credit good. The black economy is embodied in the Italian cash good; market transactions go unrecorded and untaxed. On the right-hand side, we have the gross return on bonds purchased last period, plus receipts from last period's sales, minus payments for credit goods invoiced last period.

German households face a similar budget constraint:

$$(4^*)\quad m_t^* + b_{It}^*/e_t + b_{Gt}^* + \tau_t^* p_t^* n \le R_{t-1}b_{It-1}^*/e_t + R_{t-1}^* b_{Gt-1}^* + p_{t-1}^*(n - y_{t-1}^* - q_{t-1}^* z_{t-1}^*).$$

[7] See Lucas (1987) or Lucas and Stokey (1987) and the references therein for a more detailed description of the framework outlined in this paragraph. We have extended their structure to a two-country setting.

[8] Here we are applying Elhanan Helpman and Assaf Razin's (1984) "buyer's system" to the invoicing of credit goods. Some restriction on the use of currencies is required if exchange rates are to be determinate; this is a reflection of the fact that there is no deep theory of currency substitution embodied in the cash in advance paradigm.

q^* ($\equiv p_t^{z*}/p_t^*$) is the relative price of foreign credit goods; note that $q_t = c^2/q_t^*$. The only difference between (4) and (4*) is the tax structure. There is no black economy in Germany. The German government levies a tax, τ_t^*, on all German value added.

In deriving these household constraints, we have assumed that the cash in advance constraints were binding in the previous period; that is,

$$(5) \qquad m_t = p_t x_t \quad \text{and} \quad m_t^* = p_t^* x_t^*.$$

If R_t and R_t^* are greater than one, then money is a dominated asset, and the constraints will indeed be binding. In one of the equilibria we derive, R_t^* is equal to one, and German investors are indifferent between holding money and bonds; in this case, we assume (without loss of generality) that (5) still holds and savings go into bonds.

Households maximize utility subject to budget constraints. Their first-order conditions are

$$(6) \qquad \frac{\partial u/\partial y_t}{\partial u/\partial x_t} = R_t^{-1} + \theta_t,$$

$$\frac{\partial u/\partial z_t}{\partial u/\partial x_t} = q_t\left(R_t^{-1} + \theta_t\right),$$

$$\frac{\partial u/\partial x_t}{\partial u/\partial x_{t+1}}(1+\delta) = R_t\left[\frac{p_t}{p_{t+1}}\right],$$

$$\frac{\partial u/\partial y_t^*}{\partial u/\partial x_t^*} = R_t^{*-1},$$

$$\frac{\partial u/\partial z_t^*}{\partial u/\partial x_t^*} = q_t^* R_t^{*-1},$$

$$\frac{\partial u/\partial x_t^*}{\partial u/\partial x_{t+1}^*}(1+\delta) = R_t^*\left[\frac{p_t^*}{p_{t+1}^*}\right].$$

Marginal rates of subscription are set equal to relative prices.

Governments also face budget constraints in the financial exchange:

$$(7) \qquad R_{t-1}(b_{It-1} + b_{It-1}^*) + p_{t-1}(g - T) \le b_{It}^* + b_{It} + m_t - m_{t-1} + \theta_t p_t(y_t + q_t z_t)$$

$$R_{t-1}^*(b_{Gt-1} + b_{Gt-1}^*) + p_{t-1}^*(g + q_{t-1}^* T) \le b_{Gt} + b_{Gt}^* + m_t^* - m_{t-1}^* + \tau_t^* p_t^* n.$$

On the left-hand side of each of these constraints, the first term is principal plus interest on the debt, while the second is a settling of last period's credit transactions. T is a transfer from the German government to the Italian government; we will consider combined EC policies that make such transfers possible. Governments are credit consumers; g and g^* are bought on credit. We also assume that intergovernmental transfers are credits that are settled in the next financial exchange.[9] On the right-hand side of the constraints, the first terms are new debt issue, and the last is tax proceeds.

We let h_t and h_t^* be the growth rates of the two monies.

$$(8) \qquad h_t \equiv (m_t - m_{t-1})/m_t$$

and

$$h_t^* \equiv (m_t^* - m_{t-1}^*)/m_t^*.$$

Seignorage collections can be expressed as

$$(9) \qquad m_t - m_{t-1} = h_t m_t$$

and

$$m_t^* - m_{t-1}^* = h_t^* m_t^*;$$

[9] More precisely, T is in units of the Italian cash good; so, the mark value of the credit for last period's transfer in the current financial exchange is $P_{t-1}T/e_t$. Governments incur the same valuation and currency conversion costs as the private sector; so, the total cost of the transfer is $cp_{t-1}T/e_t$ in marks or $c(p_{t-1}T/e_t p_{t-1}^*)(p_{t-1}^*/p_t^*) = q_{t-1}^*(p_{t-1}^*/p_t^*)$ in units of the German cash good.

so, h_t and h_t^* may be viewed as seignorage tax rates.

In this paper, we only consider stationary equilibria.[10] A number of results for stationary equilibria follow immediately from the arbitrage conditions, (3), the cash in advance constraints, (5), and the first-order conditions, (6):

$$(10)\qquad \begin{aligned} h_t &= \pi_t \equiv (p_t - p_{t-1})/p_t, \\ h_t^* &= \pi_t^* \equiv (p_t^* - p_{t-1}^*)/p_t^*, \\ 1-\hat{e}_{t+1} &= (1-\pi_t)/(1-\pi_t^*), \\ \hat{e}_t &\equiv (e_t - e_{t-1})/e_t, \\ R_t^{-1}(1+\delta) &= m_t/m_{t+1} = 1-h_{t+1}, \\ R_t^{*-1}(1+\delta) &= m_t^*/m_{t+1}^* = 1-h_{t+1}^*. \end{aligned}$$

These results are standard. Inflation rates are equal to money growth rates. The rate of depreciation is approximately equal to the difference in inflation rates, and nominal interest rates increase with money growth rates.

In a stationary equilibrium, the first-order conditions become

$$(11)\qquad \frac{\partial u/\partial y}{\partial u/\partial x} = \frac{1-h}{1+\delta} + \theta,$$

$$\frac{\partial u/\partial z}{\partial u/\partial x} = q\left[\frac{1-h}{1+\delta} + \theta\right],$$

$$\frac{\partial u/\partial y^*}{\partial u/\partial x^*} = \frac{1-h^*}{1+\delta}$$

and

$$\frac{\partial u/\partial z^*}{\partial u/\partial x^*} = q^*\left[\frac{1-h^*}{1+\delta}\right].$$

The household budget constraints become

$$(12)\qquad \begin{aligned} x + \theta(y+qz) &= (1-h) \\ &\quad \times (n-y-qz) \\ x^* + \tau^* n &= (1-h^*) \\ &\quad \times (n - y^* - q^* z^*), \end{aligned}$$

and the government budget constraints become[11]

$$(13)\qquad (1-h)(g-T) = \theta(y+qz) + hx$$

and

$$(1-h^*)(g+q^*T) = \tau^* n + h^* x^*.$$

In addition, there is no borrowing or lending in equilibrium, so trade balances must equal the intergovernmental transfer. For example, the Italian trade deficit, in units of the Italian cash good, is equal to[12]

$$(14)\qquad qz - cz^* = T.$$

The EC policymaker's constraint is found by aggregating the Italian and German gov-

[10] Given our assumptions, there is no reason to consider nonstationary solutions. Government spending is constant over time; so there is no need for intertemporal tax spreading. There would be issues of time consistency with the monetary policies we consider, but we have assumed that policymakers are precommitted to cooperative policies.

[11] If governments were cash customers, instead of credit customers, the budget constraints would be

$$g = \theta(y+qz) + h(x+g)$$

and

$$g^* = \tau^* n^* + h^*(x^* + g^*),$$

which are equivalent to (13).

Some governments try to increase the base for their seignorage tax by making the private sector use domestic currency for government transactions. In the present framework, such efforts are pointless. If g is a cash good, the seignorage revenue goes up by hg, but if g is bought on credit, payments are inflated away by a factor $\pi g = hg$.

[12] To verify (14), use the Italian government's budget constraint to eliminate $\theta(y+qz)$ in the Italian household's budget constraint, and subtract the result from the Italian resource constraint. Equivalent operations with the German constraints give the German trade surplus, $cz - q^*z^* = q^*T$, which is equal to the Italian deficit since $q = c^2/q^*$.

ernment budget constraints:

$$(15)\quad g+g/q^*=[\theta(y+qz)+hx]/(1-h) + (\tau^*n+h^*x^*)/q^*(1-h^*).$$

T, the intergovernmental transfer, is now implicit. Finally, it should be noted that not all of these budget constraints are independent; using the resource constraints and the trade deficit equation, the household budget constraints, (12), imply the EC policymaker's constraint, (15).

Summarizing, a *stationary equilibrium in the two currency version of the model* is defined by the two resource constraints, (2), the four first-order conditions, (11), and the two household budget constraints, (12). If an equilibrium exists, the variables $x, y, z, x^*, y^*, z^*, q$, and the tax rates h, θ, h^*, τ^* must satisfy these eight equations.

Two modifications must be made in the model if Europe adopts a common currency, the ECU. First, since there are no longer any valuation and currency conversion costs, $c=1$. Second, the cash in advance constraints, (5), reduce to a single equation,

$$(16)\quad m_t=p_tx_t+p_t^*x_t^*=p_t(x_t+q_tx_t^*) = p_t^*(x_t/q_t+x_t^*),$$

where m_t is the supply of ECU's.

In a stationary equilibrium,

$$(17)\qquad \pi=\pi^*=H,$$

where H is the growth rate of ECU's. The household first-order conditions, (11), become

$$(18)\quad \frac{\partial u/\partial y}{\partial u/\partial x}=\frac{1-H}{1+\delta}+\theta,$$

$$\frac{\partial u/\partial z}{\partial u/\partial y}=q\left[\frac{1-H}{1+\delta}+\theta\right],$$

$$\frac{\partial u/\partial y^*}{\partial u/\partial x^*}=\frac{1-H}{1+\delta}$$

and

$$\frac{\partial u/\partial z^*}{\partial u/\partial y^*}=q^*\left[\frac{1-H}{1+\delta}\right].$$

The household budget constraints, (12), become

$$(19)\quad x+\theta(y+qz)=(1-H)\times(n-y-qz)$$

$$x^*+\tau^*n=(1-H)\times(n-y^*-z^*/q),$$

and the EC policymaker's constraint, (15), is replaced by

$$(20)\quad (1-H)(g+qg) =\theta(y+qz) +q\tau^*n+H(x+qx^*).$$

Once again, the policymaker's budget constraint is redundant; using the resource constraints, it can be derived from the household budget constraints.

Summarizing, a *stationary equilibrium in the single currency version of the model* is defined by the two resource constraints, (2), the four first-order conditions, (18), and the two budget constraints, (19). If an equilibrium exists, the variables $x, y, z, x^*, y^*, z^*, q$, and the tax rates H, θ, and τ^* must satisfy these eight equations.

II. The Optimal Currency Area Problem

If the EC keeps its two currencies, then there are four tax rates (h, h^*, θ, and τ^*) to be set. The seignorage tax distorts the decision between cash and credit goods; so does the Italian VAT, since it falls only on credit goods. The German VAT is not distortionary because it falls symmetrically on cash and credit goods. Having four separate taxes helps in spreading tax distortions, but the multiplicity of currencies imposes extra costs on producers who sell abroad. If the EC adopts a common currency, these extra costs disappear, but there will only be three tax rates (H, θ, and τ^*) available for tax spreading. The EC will be an optimal currency area if valuation and currency conversion costs are more important than tax spreading.

To be more precise, we assume that the EC policymaker chooses the number of currencies (as well as the tax rates) to maximize a weighted sum of the utilities of Italian and German households. Since Italian and German households have the same utility functions, labor endowments and production possibilities, weights are chosen to ensure that $U = U^*$. (A more appealing approach might have been to use the Nash Bargaining solution. However, since it was beyond the scope of the present paper to model a noncooperative game between Italy and Germany, there was no obvious way of finding a status quo point.)

First, we calculate the best that can be done if multiple currencies are retained; that is, we solve the problem

$$(21)\quad \max \mathrm{W}^{mc} \equiv wU + (1-w)U^* \qquad \{x, y, z, x^*, y^*; z^*, q, h, h^*, \theta, \tau^*, w\}$$

subject to the conditions that define equilibrium in the two currency version of the model and the requirement that U = U*.

Then, we calculate the best that can be done if a common currency is adopted; that is, we solve the problem

$$(22)\quad \max W^{cc} \equiv wU + (1-w)U^* \qquad \{x, y, z, x^*, y^*, z^*, q, H, \theta, \tau^*, w\}$$

subject to the conditions that define equilibrium in the one currency version of the model and the requirement that $U = U^*$.

Finally, we compare the results of these two problems. W^{cc} is the utility each country obtains under a common currency, and W^{mc} is the utility each country obtains with multiple currencies. If W^{cc} is greater than W^{mc}, then the EC is an optimal currency area.

III. Tax Spreading with Two Currencies

Proposition 1 states that with two currencies there are enough tax instruments available to eliminate all of the tax distortions and obtain an efficient solution.

PROPOSITION 1: *The solution to problem (21) is also the solution to the social planner's problem,*

$$\mathrm{Max}\, W = (1/2)U + (1/2)U^* \qquad \{x, y, z, x^*, y^*, z^*\}$$

subject to the resource constraints, (2).

The social planner, unlike the EC policymaker, can allocate goods and services without going through markets or satisfying cash in advance constraints. U is equal to U^* in the planner's solution because of the symmetry in our model. In addition, the planner's allocation is the best that can be achieved given the resource constraints, (2); that is, $W^{mc} \leq W$. To prove Proposition 1, it suffices to show that the EC policymaker can find tax rates that make a market economy support the planner's allocation.

The first-order conditions for the planner's problem are

$$(23)\qquad \frac{\partial u/\partial y}{\partial u/\partial x} = \frac{\partial u/\partial y^*}{\partial u/\partial x^*} = 1$$

and

$$\frac{\partial u/\partial z^*}{\partial u/\partial x} = \frac{\partial u/\partial z}{\partial u/\partial x^*} = c.$$

In the planner's solution, marginal rates of substitution in consumption are set equal to marginal rates of transformation in production. One unit of labor produces one cash good or one home credit good, but c units of labor are required for a credit good consumed abroad. So, the marginal rate of substitution between cash goods and home credit goods is equal to one, and the marginal rate of substitution (in the planner's welfare function) between cash goods and credit goods consumed abroad is equal to c. Tax rates must be chosen so that after tax relative prices reflect these tradeoffs.

First, consider the margin between cash goods and home credit goods. The marginal rate of substitution between these goods must equal one. The tax rates that achieve this relative price can be deduced from the household first-order conditions, (6) or (11). R^*, the gross nominal return on German bonds, must be equal to one to avoid distortions between the German cash and credit goods. A consumer has to hold cash instead of bonds to buy a cash good, but this causes no distortion if money and bonds bring the same return. To make R^* equal one, the German seignorage tax must be set in accordance with Friedman's rule; that is,

$$h^* = -\delta. \tag{24}$$

In Italy, the story is a little more complicated. The VAT can also cause a distortion between cash and home credit goods because the black market escapes taxation; θ and h must be set so that

$$R^{-1} + \theta = \frac{1-h}{1+\delta} + \theta = 1. \tag{25}$$

The seignorage tax on cash goods balances the VAT on credit goods, and the relative price is unaffected.

Next, consider the margin between cash goods and credit goods sold abroad. The marginal rate of substitution (in the planner's welfare function) between cash goods and credit goods consumed abroad must equal c. However, by the symmetry of the planner's solution,

$$\frac{\partial u/\partial z^*}{\partial u/\partial x} = \frac{\partial u/\partial z}{\partial u/\partial x} \tag{26}$$

and

$$\frac{\partial u/\partial z}{\partial u/\partial x^*} = \frac{\partial u/\partial z^*}{\partial u/\partial x^*}.$$

The tax rates that make the marginal rate of substitution between cash goods and imported credit goods equal to c can again be inferred from the household first-order conditions. If tax rates satisfy restrictions (24) and (25), and if $q = q^* = c$, then this margin will also be met.

The planner's solution is symmetric. So, z is equal to z^*, and (14) implies that q (and thus q^*) is equal to c if the intergovernmental transfer, T, is equal to zero.

Summarizing, if tax rates satisfy restrictions (24) and (25) and if the intergovernmental transfer is zero, then the first-order conditions for the planner's problem are satisfied. Finally, tax revenues must finance government spending in each country. So, the tax rates must also satisfy the government budget constraints, (13). This will be possible if (as we assume) g and δ are not too big. This completes the proof of Proposition 1.

Corollary 1 emphasizes the fact that optimal tax spreading requires higher inflation in Italy than in Germany, and therefore a continuous depreciation of the lira. Any attempt to limit this depreciation would be counterproductive; tax distortions would be introduced, and welfare would be lower.

COROLLARY 1: *In the optimal multiple currency solution*, $\pi^* = -\delta$, $\pi > \pi^*$, *and* $\hat{e} \approx \pi - \pi^* = (1+\delta)\theta > 0$.

The corollary follows from equation (10) and restrictions (24) and (25).

IV. When Is the EC an Optimal Currency Area?

If the EC adopts a common currency, then the valuation and currency conversion costs are eliminated; c reverts to unity. However, since there is also one less tax rate for the EC policymaker to set, there are too few instruments to eliminate tax distortions entirely. One might expect that the answer hinges on the size of c: if c is small, then the EC should retain its separate currencies to facilitate tax spreading; if c is large, then the EC is an optimal currency area.

More specifically, one might expect something like Figure 1. W^{mc} (welfare in the multicurrency solution) should be a decreasing function of c. When c is equal to 1, multiple currencies impose no extra costs. So, tax spreading advantages should make multiple currencies dominate a common currency; that is, $W^{mc}(1)$ should be greater than

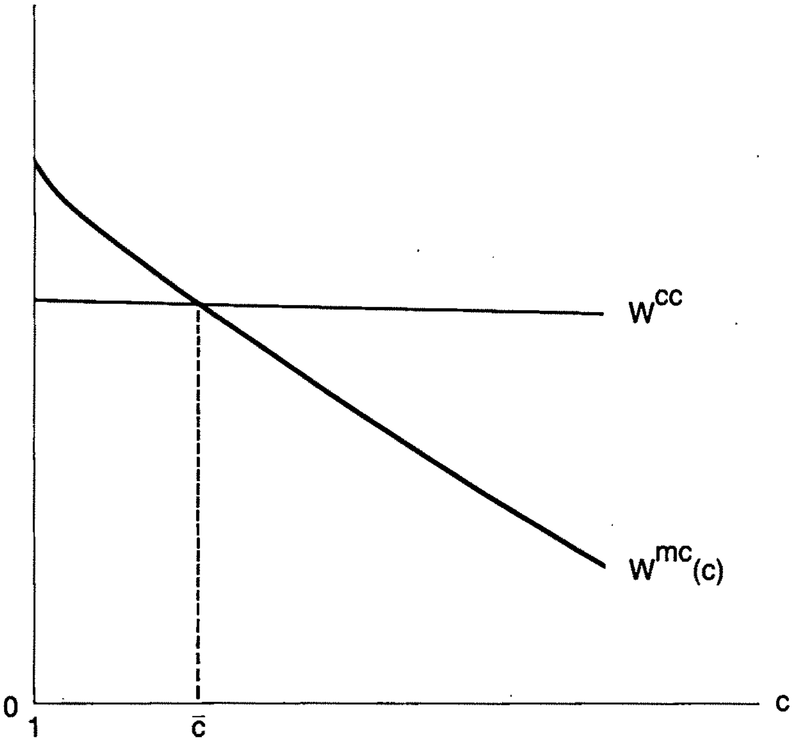


FIGURE 1

W^{cc} (welfare in the common currency solution). However, as c increases, the attractiveness of multiple currencies should diminish. For c greater than some critical value, $\bar{c}$, the EC should be an optimal currency area.

Lemma 1 states that welfare under multiple currencies is indeed a decreasing function of c, and that the EC cannot be an optimal currency area for values of c close to one.

LEMMA 1: W^{mc} *is a function of* c, $dW^{mc}(c)/dc < 0$, *and* $W^{mc}(1) > W^{cc}$.

The planner's problem described in Proposition 1 has a unique interior solution and W^{mc} is a differentiable function of c because of the restrictions placed on $u(x, y, z)$. W^{mc} is a decreasing function of c because $dW^{mc}(c)/dc = -\lambda z^* - \lambda^* z$, where λ and λ^* are the positive Lagrange multipliers in the planner's problem.[13] $W^{mc}(1) \geq W^{cc}$ since the planner's solution cannot be improved upon. $W^{mc}(1) > W^{cc}$ follows from the fact that the planner's solution cannot be duplicated in the common currency version of the model.

What remains to be shown in Figure 1 is that $W^{mc}(c) < W^{cc}$ for large values of c. It turns out that this need not be the case, even as c goes to infinity. This seemingly paradoxical result can be explained as follows. As c rises, international trade is driven out by the high cost of selling to foreigners. However, if home goods and imports are very close substitutes, then consumers lose little utility in switching to home goods, and the policy-

[13] This fact follows from the symmetry of the planner's problem and the envelope theorem; see, for example, Appendix A.11 in Hal Varian (1978).

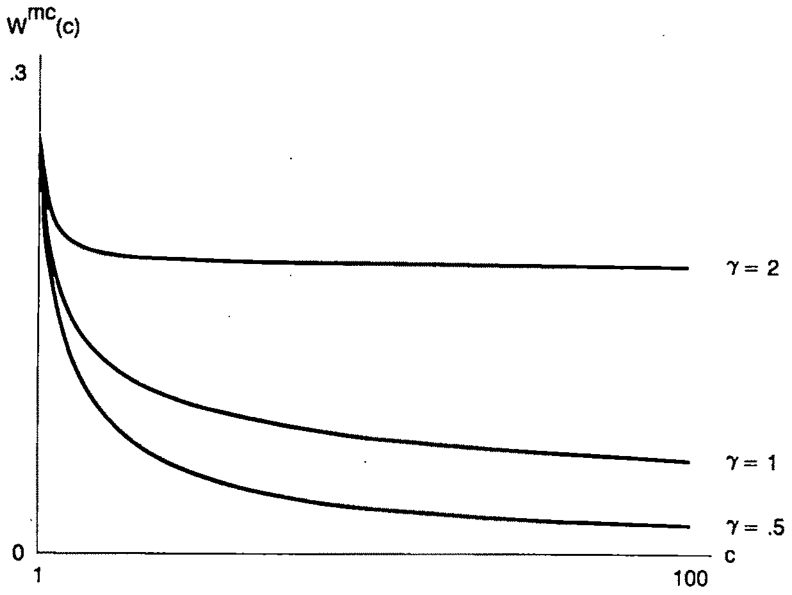


FIGURE 2

maker can continue to take advantage of the tax spreading opportunities afforded by a multiplicity of currencies. If on the other hand home goods and imports are not close substitutes, then consumers do lose utility in the switch from imports to home goods, and $W^{mc}(c) < W^{cc}$ for large values of c.

It is convenient at this point to restrict our attention to the class of CES (constant elasticity of substitution) utility functions; let

$$(27)\quad u(x,y,z) = \left[\mu_1 x^{1-1/\gamma} + \mu_2 y^{1-1/\gamma} + \mu_3 z^{1-1/\gamma}\right]^{(1-1/\gamma)^{-1}},$$

where $\mu_1+\mu_2+\mu_3=1$ and $0<\gamma<\infty$; γ is the elasticity of substitution.

Lemma 2 describes $W^{mc}(c)$ for this class of utility functions.

LEMMA 2: *If utility is CES, then*

$$W^{mc}(c) = \begin{cases} (n-g) \\ \quad \times\left[\mu_1^\gamma + \mu_2^\gamma + \mu_3^\gamma c^{(1-\gamma)}\right]^{(\gamma-1)^{-1}} \\ \qquad\qquad\qquad \textit{for } \gamma \neq 1 \\ (n-g)\,\mu_1^{\mu_1}\mu_2^{\mu_2}\mu_3^{\mu_3}c^{-\mu_3} \\ \qquad\qquad\qquad \textit{for } \gamma = 1 \end{cases}$$

and as $c \to \infty$,

$$W^{mc}(\infty) = \begin{cases} 0 \\ \qquad\qquad \text{for } 0<\gamma\le 1 \\ (n-g)\left[\mu_1^\gamma + \mu_2^\gamma\right]^{(\gamma-1)^{-1}} \\ \qquad\qquad \text{for } 1<\gamma<\infty. \end{cases}$$

Lemma 2 is proved by solving the planner's problem with CES utility functions. Figure 2 illustrates $W^{mc}(c)$ for various values of γ. If γ is less than or equal to one, W^{mc} goes to zero as c gets very large.

Proposition 2 states our basic result. The EC will be an optimal currency area if valuation and currency conversion costs are big enough and if the elasticity of substitution is not too big.

PROPOSITION 2: *If utility is* CES, *then*

i. *if the elasticity of substitution is less than or equal to one, there exists a critical value, $\bar{c}$, of the valuation and currency conversion costs such that $W^{mc}(c) > W^{cc}$ if $c < \bar{c}$ and $W^{mc}(c) < W^{cc}$ if $c > \bar{c}$.*

ii. *if the elasticity of substitution is high, then a critical value, $\bar{c}$, need not exist; $W^{mc}(c)$ may be greater than W^{cc} for all values of c.*

TABLE 3—THE IMPORTANCE OF γ

γ	$W^{mc}(\infty)$	W^{cc}
1.0	0.000	0.271
5.5	0.291	0.291
10.0	0.313	0.307

To prove the first part of Proposition 2, it suffices to show that $W^{cc} > 0$, because Lemma 2 implies that W^{mc} goes to zero as c goes to infinity and Lemma 1 states that $W^{mc}(1) > W^{cc}$. The restrictions on $u(x, y, z)$ guarantee that the consumer's problem has an interior solution. Therefore, if (as we assume) problem (22) has a solution, it is an interior solution. The CES utility function gives strictly positive utility if all of its arguments are positive.

To prove the second part of Proposition 2, it suffices to find an example where $W^{mc}(\infty) > W^{cc}$. In Table 3, we report a set of numerical examples in which $\delta = 0.05$, $n = 1$, $g = 0.2$, $\mu_1/\mu_2 = 1.5$ and $\mu_2 = \mu_3$. When $\gamma = 1$, W^{mc} goes to zero as c goes to infinity; W^{mc} is less than W^{cc} for large values of c. When $\gamma = 10$, W^{mc} goes to 0.313 as c goes to infinity, while W^{cc} is equal to 0.307; W^{mc} is greater than W^{cc} for all values of c.

V. Some Illustrative Examples

The "simple" model outlined in Section I is really rather complex, and this complexity makes it difficult to say more about the factors that determine the size of the critical value, $\bar{c}$. In this section, we suggest some propositions about the size of $\bar{c}$ based on two sets of numerical examples.

The examples all use a Cobb-Douglas utility function

$$(28) \qquad u(x, y, z) = x^{\mu_1} y^{\mu_2} z^{\mu_3},$$

where $\mu_1 + \mu_2 + \mu_3 = 1$. This is of course the CES utility function with $\gamma = 1$. And in all of the examples, $\delta = 0.05$ and $n = 1$.

The first set of examples illustrates how the critical value of c might depend on the size of government spending. $\mu_1 = \mu_2 = \mu_3 = 1/3$ and g is allowed to vary between 0 and 0.4. The results are reported in Table 4; they suggest that the critical value of c is an increasing function of g. This is a plausible proposition. As g increases, more revenue must be raised. Tax spreading becomes more important, and the value of c for which the EC becomes an optimal currency area is higher.

TABLE 4—$\bar{c}$ AND THE SIZE OF g

g	0.0	0.1	0.2	0.3	0.4
$\bar{c}$	1.00002	1.00171	1.00667	1.01623	1.03265

TABLE 5—$\bar{c}$ AND OPENNESS

μ_3/μ_1	0.5	0.75	1.0	1.25	1.5
$\bar{c}$	1.0134	1.0089	1.0067	1.0053	1.0045

The surprising result in Table 4 may be the low numbers obtained for the critical value of c. Even if the government consumes as much as twenty percent of output, valuation and currency conversion costs of only seven-tenths of one percent of the cost of producing the traded good make the EC an optimal currency area. The critical value does rise rapidly as g is increased.

The second set of examples illustrates the effect of openness on the critical value of c. We measure openness by the relative importance of imported goods in the utility function; that is, we let $\mu_1 = \mu_2$, and we take the ratio μ_3/μ_1 as our measure of openness. $g = 0.2$ and μ_3/μ_1 is allowed to vary between 0.5 and 1.5. The results are reported in Table 5; they suggest that the critical value of c goes down as openness is increased. This too is a plausible proposition.[14] As imports become more valuable, lowering the cost of producing them becomes more important.

Summarizing, the examples presented in this section suggest the following propositions: (1) Small valuation and currency conversion costs can make the EC an optimal

[14] Mundell (1961) also found that openness was a factor leading countries to form a currency union.

currency area. (2) High levels of government spending make the EC less likely to be an optimal currency area. (3) Openness makes the EC more likely to be an optimal currency area.

V. Limitations and Suggestions for Future Research

Throughout this paper, we have assumed that the Italian black economy is a cash good, subject to a lira denominated cash in advance constraint. In Section III, this allowed the EC policymaker to tax the black market separately, using the Italian inflation tax. Alan Stockman has pointed out that there is an obvious incentive here for black market participants to switch to marks, which are subject to a lower inflation tax; this possibility is simply ruled out by the cash in advance constraints in our model. It would therefore be interesting to replace our cash in advance constraints with a restriction that allowed currency substitution. Would Stockman's concern constitute an argument for currency controls? We do not know.

Throughout this paper, we have also assumed that all fiscal policy decisions are made by a central EC policymaker. Present proposals for European economic integration do not go nearly so far. It would be interesting to introduce Italian and German policymakers who play a Nash game with each other in their attempts to maximize their own citizens' welfare. In a multicurrency setting, the "1992" proposals would be represented by set of rules of the game. For example, protective tariffs would be outlawed. In a common currency setting, we might envision a central EC policymaker who issues the currency and distributes the seignorage.[15] The Italian and German policymakers would be free to set the remaining tax rates.

Finally, we have also assumed that policymakers can precommit to monetary policies that may be time inconsistent. (The policies described in Section III do happen to be time consistent, since there are no distortions in the equilibrium that results; this would not be the case if, say, a labor-leisure decision were added.) We noted in the introduction that the EMS is sometimes justified on the grounds of credibility; Italy is thought to be able to solve its precommitment problem by tying the lira to the mark.[16] It would therefore be interesting to drop our assumption about precommitment and investigate time-consistent solutions.[17] We speculate that Italy would have the higher inflation rate, and an inflation bias. Italy may wish to tie the lira to the mark to reduce the inflation bias, even if this interferes with optimal tax spreading. In other words, credibility may have to be weighed along with the valuation and currency conversion costs when defining an optimal currency area.

[15] Alessandra Casella and Jonathan Feinstein (1988) demonstrate that difficulties arise if both the Italian and German policymakers are allowed to issue the common currency.

[16] Canzoneri and Henderson (1988) question whether a credibility problem can be so easily solved: why is an announcement about an exchange rate policy more binding than an announcement about a money supply policy?

[17] To do so, we would have to introduce a cost to actual (as opposed to anticipated) inflation; see Herschel Grossman (1988) or Maurice Obstfeld (1988).

REFERENCES

Bailey, Martin and Tavlas, George, "Trade and Investment Performance Under Floating Exchange Rates: The U.S. Experience," IMF Working Paper, May 1988.

Barro, Robert, "Interest-Rate Targeting," *Journal of Monetary Economics*, January 1989, *23*, No. 1, 3–30.

______ **and Gordon, David,** "Rules, Discretion and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics*, 1983, *12*, 101–21.

Canzoneri, Matthew and Gray, Jo Anna, "Monetary Policy Games and the Consequences of Noncooperative Behavior," *International Economic Review*, October 1985, *26*, No. 3, 547–63.

Canzoneri, Matthew and Henderson, Dale, "Is Sovereign Policymaking Bad?" *Carnegie-Rochester Conference Series on Public Policy*, 1988, *28*, 93–140.

Casella, Alessandra and Feinstein, Jonathan, "Management of a Common Currency," NBER Working Paper, October 1988.

Chang, Roberto, "Does International Coordination of Fiscal Deficits Matter?" New York University Working Paper, November 1988.

Contini, Bruno, "The Second Economy of Italy," in Vito Tanzi, ed., *The Underground Economy in the United States and Abroad*, Lexington, MA: Lexington Books, 1982.

de Grazia, Raffaele, *Clandestine Employment*, International Labor Organization, Geneva, 1984.

Delors, Jacques, "Report on Economic and Monetary Union in the European Community," Coauthor-Publisher, Committee for Study of Economic and Monetary Union, prepared in response to mandate of European council, April 1989.

Dornbusch, Rudiger, (1988a) "Money and Finance in European Integration," EFTA Seminar, Geneva, 27 January 1988.

______, (1988b) "The European Monetary System, the Dollar, and the Yen," in Giavazzi, Francesco et al., eds., *The European Monetary System*, Cambridge, MA: Cambridge University Press, 1988.

Drazen, Allan, "Monetary Policy, Capital Controls and Seignorage in an Open Economy," University of Pennsylvania Working Paper, October 1988.

Fischer, Stanley, "Seignorage and the Case for a National Money," *Journal of Political Economy*, 1982, *90,* No. 21, 295–313.

Frenkel, Jacob and Razin, Assaf, "Exchange-Rate Management Viewed as Tax Policies," NBER Working Paper No. 2653, July 1988.

Genberg, Hans, "Exchange Rate Management and Macroeconomic Policy: A National Perspective," The Graduate Institute of International Studies, July 1988.

Giavazzi, Francesco, "The Exchange Rate Question in Europe," University of Bologna Working Paper, February 1988.

Grilli, Vittorio, "Seignorage in Europe," in Marcello de Cecco and Alberto Giovannini, eds., *A European Central Bank?* Cambridge: Cambridge University Press, 1989.

Grossman, Herschel, "Inflation and Reputation with Generic Policy Preferences," unpublished manuscript, revised September 1988.

Helpman, Elhanan and Razin, Assaf, "The Role of Saving and Investment in Exchange Rate Determination Under Alternative Monetary Mechanisms," *Journal of Monetary Economics*, May 1984, *13*, 307–25.

Kehoe, Patrick, "Policy Cooperation Among Benevolent Governments May Be Undesirable," FR Bank of Minneapolis Working Paper No. 373, October 1987.

Langefeldt, Enno, "Is a Growing Unobserved Sector Undermining Monetary Policy in the Federal Republic of Germany?" in Wulf Gaertner and Alois Wenig, eds., *The Economics of the Shadow Economy, Proceedings of the Conference on the Economics of the Shadow Economy*, University of Bielefeld, October 1983.

Lucas, Robert, "Principles of Fiscal and Monetary Policy," *Journal of Monetary Economics*, January 1986, *17*, 117–34.

______ **and Stokey, Nancy,** "Money and Interest in a Cash-in-Advance Economy," *Econometrica*, May 1987, *55,* No. 3, 491–513.

Mankiw, N. Gregory, "The Optimal Collection of Seignorage: Theory and Evidence," *Journal of Monetary Economics*, September 1987, *20*, 327–41.

Mundell, Robert, "A Theory of Optimal Currency Areas," *American Economic Review*, September 1961, *51*, 657–65.

Obstfeld, Maurice, "Dynamic Seignorage Theory: An Exploration," unpublished manuscript, July 1988.

Poole, William, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics*, May 1970, *84*, 197–216.

Varian, Hal, *Microeconomic Analysis*, New York: Norton & Co., 1978.

Vegh, Carlos and Guidotti, Pablo, "Optimal Taxation Policies in the EMS: A Two Country Model of Public Finance," IMF Working Paper, December 1988.

Foreign Exchange Rate Expectations: Micro Survey Data

By TAKATOSHI ITO*

This paper analyzes the panel data of biweekly surveys on the yen/dollar exchange rate expectations of forty-four institutions for two years, and contains four major findings. First, market participants are heterogeneous; that is, significant "individual effects" exist in their expectation formation. Second, the individual effects have a characteristic of "wishful expectations": exporters expect yen depreciation, and importers expect yen appreciation (relative to others). Third, many violate the rational expectations hypothesis. Fourth, forecasts with long horizons showed less yen appreciation than those with short horizons. Cross-equation constraints implied by the consistency of the forecast term structure are strongly rejected in the data.

As rational expectations have become a popular benchmark for thinking about financial and macroeconomic hypotheses, many economists have become more interested in directly measuring the expectations of market participants. Although survey data for many domestic variables, including interest rates and inflation rates, have been frequently analyzed by many investigators (see, for example, Frederic S. Mishkin, 1983, ch. 4), it is only recently that survey data on foreign exchange rates have become available and been analyzed. Kathryn M. Dominguez (1986) and Jeffrey A. Frankel and Kenneth A. Froot (1987a,b) have exploited the survey data made available by the Money Market Service (MMS), the Amex Bank Review and the Economist Financial Report.[1]

The surveys that were investigated by Dominguez and by Frankel and Froot have had only their median responses reported. Heterogeneity among market participants, if it exists, has been aggregated out. If the market consists of homogeneous agents that share the same forecasting model with common beliefs (priors) and information, then the median response would sufficiently describe the market in terms of forecasts. However, if market participants differ in the fore-

*Institute of Economic Research, Hitotsubashi University, Tokyo 186, Japan; and National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138. I gratefully acknowledge the help and encouragement from Tomomitsu Oba, Shoji Ochi, and Eisuke Sakakibara in allowing me access to the JCIF proprietary data for academic purposes only. Discussions with Jeffrey A. Frankel, Kenneth A. Froot, Hidehiko Ichimura, Mark Machina, Maurice Obstfeld, Thomas Sargent, and Christopher Sims were very useful. I have benefited from comments received from the participants of a session in the Econometric Society meeting of December 1988, and of seminars at Harvard University, the University of Minnesota, the University of California at San Diego, the National Bureau of Economic Research, and the Bank of Japan. Comments by three anonymous referees on the earlier version of the paper contributed to several significant improvements. Financial support from the National Science Foundation (SES-88-08828) is gratefully acknowledged.

[1] Dominguez (1986) used the Money Market Service (MMS) data from 1983 to 1985 to test a rational expectations hypothesis. Unbiasedness and the independence of forecast errors from the forward premium were tested. She found that survey forecasts were no better than the spot rate in predictive power and that rationality was in general rejected. In addition to the MMS data, Frankel and Froot (1987a,b) exploited the survey data collected by Amex Financial Service and also the Economist, which have longer sample periods and different forecast horizons. They found that expectations do respond to exchange rate changes. Moreover, short-term forecasts are more "destabilizing" than long-term forecasts; that is, the response to the degree of forecasted appreciation in response to appreciation is larger in the short-term horizon than in the long-term one.

casting characteristics, then focusing on the median misses the most interesting questions, such as whether the differences persist or are temporary, whether the differences are correlated with the participant's traits, and whether a rationality hypothesis is more likely to be rejected in individual data. Only individual responses of survey data can answer these questions.

In this paper, I will use the survey data collected by the Japan Center for International Finance (JCIF) in Tokyo, which allows me to investigate the individual responses in the survey. In particular the JCIF data set has two distinct advantages over the data used by Dominguez and by Frankel and Froot. First, the JCIF data consist of individual responses with no missing observations. This is the first paper to study the individual responses of exchange rate expectations, although individual responses of inflation expectations were studied before Stephen Figlewski and Paul Wachtel (1981). Second, not only financial institutions but other companies as well are polled in the JCIF survey. Therefore, there is a chance to associate possible heterogeneity to the traits of the forecasters' industry.

There are four major findings in this paper. First, market participants are found to be heterogeneous. There are significant "individual effects" in their expectation formation. Second, the individual effects have characteristics of "wishful expectations": exporters expect a yen depreciation (relative to others), and importers expect a yen appreciation (relative to others). Third, many institutions are found to violate the rational expectations hypothesis. Most of them underestimated the degree of yen appreciation. Fourth, forecasts with long horizons showed less yen appreciation than ones with short horizons. Put differently, market participants appear to have a "bandwagon" expectation in the short run, but a "stabilizing" one in the long run. The "twist" in forecast term structure could be "consistent" (in the sense of Kenneth A. Froot and Takatoshi Ito, 1989), if an iterated substitution of a short-term forecast yields a long-term forecast. However, cross-equation constraints implied by the consistency are strongly rejected.

I. Data Summary

A. *The Data Description*

The JCIF has conducted telephone surveys twice a month, in the middle and at the end of the month, on Wednesdays, since May 1985. Forecasts of the yen/dollar exchange rate for the one-, three-, and six-month horizons are obtained from foreign exchange experts in 44 companies, including 15 banks and brokers, 4 securities companies, 6 trading companies, 9 export-oriented companies, 5 life insurance companies, and 5 import-oriented industries.[2] Each respondent is asked to give a point forecast for each horizon. In this paper, I assume that reported forecasts are the subjective means of respondents. We do not have any data on the subjective variance or range. The survey is meticulously arranged so that all 44 companies on the permanent list respond every week.

When a data set is analyzed as panel data, the mean across individuals and the mean across time should not be confused. In the following, the mean across 44 individuals at a time will be referred to as the (cross section, total) average; the mean across individuals at a time in an industry group will be referred to as the group average. The mean across time of an individual, of a group, or of the "average" will be referred to as the (time) mean of the individual, of the group, or of the average, respectively.

The JCIF calculates the total average, the standard deviation, the maximum, and the minimum of the 44 responses and also the industry group averages and the group stan-

[2] The first few surveys were conducted not on Wednesdays but on the middle and last business days of the month. However, the survey date was fixed on Wednesday after the fourth observation. A twice-a-month survey means that observations are usually biweekly, with a couple of exceptions in a year. That is, there are 24, instead of 26, observations in the JCIF data in 52 weeks. It is unfortunate that the interval is not fixed. In the following, I disregard the problem arising from a mix of two- and three-week intervals. The survey started with 42 companies and expanded to the current 44 after the fourth survey in July 1985.

TABLE 1—A. TIME MEAN OF $\{s_j^e(t,k)-s(t)\}$,[a]
MAY 1985–JUNE 1987, NUMBER OF OBSERVATIONS = 51

Horizon	1 Month	3 Month	6 Month
AVE	−1.420	−1.431	−0.044
BAN	−1.404	−1.658	−0.957
SEC	−1.097	−0.834	+0.621
TRA	−1.956	−2.453	−0.948
EXP	−0.775	−0.137	+1.736
INS	−1.746	−2.309	+0.302
IMP	−1.937	−1.536	−0.430
ACT	−2.064	−5.970	−11.987

TABLE 1—B. UNCONDITIONAL EXPECTED CHANGE, DISTRIBUTION AMONG INDIVIDUAL RESPONDENTS OF THE TIME MEAN OF FORECASTED CHANGES IN THE EXCHANGE RATE OVER THE SPECIFIED HORIZON

Horizon	1 Month	3 Month	6 Month
percent			
+5.0			
			x
			xx
		x	
			xxxx
		x	*xxxx*
	x		*x*
		xx	*xxxx*
0.0	*x*	*xx*	*xxxx*
	xx	*xxx*	*xxxx*
	xxxx	*xxxxxxxxx*	*xxxxx*
	xxxxxxxxxxxxxxx	*x*	*xxxxxx*
	xxxxxxxxxxx	*xxxxxxx*	*x*
	xxxxxxx	*xxxxxxxx*	*xx*
	xx	*xxxx*	*x*
	x	*xx*	*xx*
		x	*x*
		xx	
		x	*x*
			x
−5.5			
	Max 1.41	Max 3.25	Max 4.62
	Min −3.10	Min −4.76	Min −5.20

Mean of the (unconditional) expected changes (in percent). Not annualized or adjusted for *k*.
x = 1 respondent.

dard deviations. On the day after the survey, the JCIF informs its subscribers, including those who are polled, of the summary statistics. The total average is also released to the press and other media.

I will use, in addition to the panel data of the 44 companies, the public information part of the survey, the cross-section average (AVE) and the group averages for the different industries: banks (BAN), securities companies (SEC), trading companies (TRA), companies in the export industries (EXP), insurance companies (INS), and companies in the import industries (IMP). The unit is yen per one U.S. dollar, so that a negative movement indicates a yen "appreciation."

The spot exchange rate, $s(t)$, is measured at the closing quote in Tokyo on Wednesday of the survey week.

B. *Overview*

Table 1 shows the time means of (unconditional) expected changes (in percent) from the spot rate at the time of survey for the cross-section total average, the group averages, and (in a separate distribution table) for each individual. For the purposes of discussion, the actual (*ex post*) changes of the spot exchange rate (ACT) for each horizon are reported in the same table. For each horizon and each individual or group, subtracting the actual changes from the forecasts produces the forecast errors.

In the one-month horizon, the (total) average on a typical week showed an expected 1.4 percent yen appreciation. The group averages ranged from a 0.8 percent to a 2 percent appreciation. Relative to the total average, the export industry was the most biased toward a yen depreciation, and the trading companies and the import industries were the most biased toward a yen appreciation. Looking into individual data, one extreme predicted a 1.4 percent *depreciation* of the yen, while the other extreme predicted a 3.1 percent of appreciation. The distribution of individual forecasts has a nice unimodal distribution. The average expected appreciation of the yen in the three-month horizon was 1.4 percent, about the same as in the one-month horizon. (Note that no adjustment is made with respect to the length of horizon.)

As in the one-month horizon, the export industry shows a yen depreciation bias (from the total average), and the trading companies show a yen appreciation bias in the three-month horizon. A wide disagreement among individuals begins to appear in the three-month forecasts. It becomes a bimodal distribution: one group believes that the yen depreciates from the one-month to three-month forecast horizon, while the other believes that the yen continues to appreciate.

For the six-month horizon, the total average shows that the market expects the yen to return to nearly the prevailing level at the time of forecast. This is a sharp turnaround from the forecast of a 1.4 percent yen appreciation in three months. In fact, each of the group averages indicates that the group anticipates less yen appreciation in the six-month horizon than in either the one- or the three-month horizon.

The findings of this subsection can be summarized and related to the contents of the rest of this paper. First, the findings are highly suggestive of heterogeneous market participants. A rigorous analysis and interpretation of the heterogeneity will be provided in Section III. Second, large forecast errors were recorded during the intermittent waves of yen appreciation after September 1985. Econometric tests on various forms of the rational expectation hypothesis will be conducted in Section IV. Third, the total average and most of the group averages have a "twist" in their forecasts, a yen appreciation in the short horizon and a yen depreciation in the long horizon. Section V investigates whether such twists in expectations are internally consistent.

II. Wishful Expectations and Heterogeneity

A. *Econometric Issue — A Special Case of Panel Data*

Recall that our micro survey data set consists of 44 individuals and 51 observations. Suppose that an individual forecast formation at time t consists of a common structural part based on public information, $f(I(t))$ and an individual effect, g_j. For a given forecast horizon, k (suppressed notation), the expected exchange rate for individual j, $j=1,\dots,J$ (where in this paper $J=44$) is

$$(1)\qquad s_j^e(t)=f(I(t))+g_j+u_j(t),$$

where $s_j^e(t)$ is a k-step ahead forecast of the spot exchange rate at time t, by individual j; $u_j(t)$ is a pure random disturbance (with respect to j and t) representing, for example, a measurement or a rounding error. The cross-section average of individual forecasts,

$s^e_{AVE}(t)$ is defined as

$$(2)\quad s^e_{AVE}(t)=f(I(t))+g_{AVE}+u_{AVE}(t),$$

where $x_{AVE}(t)=(Ex_j(t))/J$; $x=s^e$, g, and u. Assume $f(I(t))$ contains a constant term so that normalization, $g_{AVE}=0$, is possible. Then subtracting each side of (2) from the corresponding side of (1), we obtain

$$(3)\quad s^e_j(t)-s^e_{AVE}(t)$$
$$=g_j+\big(u_j(t)-u_{AVE}(t)\big).$$

The estimator of the individual effect, g_j can be obtained by regressing the left-hand side of (3) on a constant over the sample period (across time). This procedure is simple and robust. It is unnecessary for the econometrician to know the exact structure of $f(I(t))$ as long as it is common to everybody for every survey date.

If the difference in the individual effects of two individuals is to be estimated, a similar method can be employed.

$$(4)\quad s^e_j(t)-s^e_h(t)$$
$$=g_j-g_h+\big(u_j(t)-u_h(t)\big),\ h\neq j.$$

A (composite) disturbance term in equations (3) and (4) has mean zero and no serial correlations if $u_j(t)$ is serially and cross-sectionally uncorrelated and $f(I(t))$ is exactly common to all individuals.

If the difference in individual beliefs extends to "idiosyncratic" coefficients on publicly available information in the structural part, $f(I(t))$, the above procedure needs to be modified but is still applicable. Suppose, for example, that the forecast is in an extrapolative form:

$$(5)\quad s^e_j(t)-s(t)$$
$$=a_j+b_{1j}\big(s(t-1)-s(t)\big)$$
$$+b_{2j}\big(s(t-2)-s(t-1)\big)+u_j(t),$$

where g_j is the difference in a_j. Then the idiosyncratic individual coefficients can be estimated by regressing the following equations, for all j:

$$(6)\quad s^e_j(t)-s^e_{AVE}(t)$$
$$=a_j-a_{AVE}$$
$$+\{b_{1j}-b_{1AVE}\}\big(s(t-1)-s(t)\big)$$
$$+\{b_{2j}-b_{2AVE}\}$$
$$\big(s(t-2)-s(t-1)\big)$$
$$+u_j(t)-u_{AVE}.$$

The above procedure parallels the technique in the panel data analysis, although in the usual examples of the panel data analysis, the right-hand-side variables take different values for different individuals. Instead, it is reasonable here to assume that the structural part and the values of regressors (i.e., the past values of the exchange rates) in exchange rate forecasts are identical for all individuals, but with possibly different coefficients.

B. *Heterogeneous Participants in the Tokyo Market*

In search of hard evidence for (or against) heterogeneity among market participants, I estimate 44 individual effects, g_j, and "group effects." In detecting the "group effect," a group average forecast calculated by the JCIF is treated as an individual j, then the total average (or another group average) is subtracted.[3]

The individual (or group) effect g_j, estimated using equation (3), are reported in Table 2.[4]

[3] Since the micro panel data set was made available on the condition that the anonymity of the source should be honored, it is impossible to aggregate the individuals into groups.

[4] For some cases, an allowance had to be made for AR(1) serial correlation in $u_j(t)-u_{AVE}(t)$, or in $u_j(t)-u_h(t)$, contrary to the assumptions mentioned earlier. This may be due to either serial correlation in u_j or deviations in $f(I(t))$ among individuals. However, many of rejection cases (i.e., confirming heterogeneity) are found without AR(1) disturbances.

TABLE 2—A. GROUP DEVIATIONS FROM THE TOTAL AVERAGE, FOR EACH HORIZON

	1 Month		3 Month		6 Month	
Horizon	*a*	*DW* or *RHO*	*a*	*DW* or *RHO*	*a*	*DW* or *RHO*
BAN	0.017	0.284	−0.228	0.530	−0.941	0.371
t-stat	(0.25)	(2.04)	(−1.28)	(4.29)	(−5.74)[a]	(2.81)
SEC	0.305	0.438	0.561	0.421	0.743	0.446
t-stat	(1.25)	(3.38)	(1.62)	(3.14)	(1.47)	(3.49)
TRA	−0.536	*DW* = 2.13	−1.022	*DW* = 1.61	−0.908	0.467
t-stat	(−4.98)[a]		(−7.56)[a]		(−2.57)[a]	(3.61)
EXP	0.645	*DW* = 2.07	1.294	*DW* = 1.62	1.832	0.435
t-stat	(8.55)[a]		(12.68)[a]		(6.11)[a]	(3.41)
INS	−0.326	0.474	−0.815	0.645	0.301	0.661
t-stat	(−1.54)	(3.72)	(−1.93)[a]	(5.86)	(0.54)	(5.99)
IMP	−0.517	*DW* = 1.47	−0.079	0.301	−0.434	0.422
t-stat	(−3.76)[a]		(−0.29)	(2.17)	(−1.39)	(3.27)

[a] It shows the "heterogeneous" group at the level of 1 percent.

TABLE 2—B. WISHFUL EXPECTATIONS, DISTRIBUTION OF INDIVIDUAL EFFECTS

		Horizon	
Horizon	1 Month	3 Month	6 Month
percent +5.0			
		x	*x*
			xx
		x	*x*
			x
	x		*xo*
		xx	*o*
	x	*xxo*	*xo*
	x		*xxoo*
	xxxoo	*oooooooooo*	*ooo*
0.0	*xoooooooooooo*	*o*	*ooo*
	xooooooooooo	*ooooooo*	*ooooo*
	xxxxooo	*oooooooo*	*ooooo*
	xxxx	*xxxo*	*xxoo*
	x	*xo*	*xo*
		x	*xo*
		xx	*x*
		x	*xx*
		x	
			o
−5.5			*xx*

x = significant individual effects;
o = insignificant individual effects.

From Table 2, panel A, we learn that for any horizon, group effects are significant for the export industry, with a depreciation bias, and for the trading companies, with an appreciation bias. A significant appreciation bias was also detected for the import industry for the one-month horizon, for the insurance industry for the three-month horizon, and for the banking sector for the six-month horizon.

The distinctive effect of exporters in contrast to importers or to trading companies can be highlighted by measuring the difference in individual effects directly, as in equa-

TABLE 3—IDIOSYNCRATIC EFFECTS: EXTRAPOLATIVE FORM

$$s^e_j(t) - s^e_{AVE}(t) = a_j - a_{AVE} + \{b_{1j} - b_{AVE}\}(s(t-1) - s(t)) + \{b_{2j} - b_{2AVE}\}(s(t-2) - s(t-1)) + u_j(t) - u_{AVE}$$

$H0$: No idiosyncratic coefficient effects, $b = 0$
(allowing for individual effect of a constant bias)
$H1$: No idiosyncratic coefficient of individual (constant) effect: $a = b = 0$

Lag	1 lag ($b_2 = 0$)						2 lag					
length	1 Month		3 Month		6 Month		1 Month		3 Month		6 Month	
Horizon	$H0$	$H1$	$H0$	$H1$	$H0$	$H1$	$H0$	$H1$	$H0$	$H1$	$H0$	$H1$
BAN F	0.122	0.103	2.51	2.48	2.34	18.19	0.433	0.338	2.60	3.13	0.732	17.3
sig	0.729	0.903	0.120	0.095	0.133	0.00*	0.651	0.798	0.086	0.035	0.487	0.00*
SEC F	0.815	1.37	0.037	1.41	0.000	1.03	0.699	1.51	0.281	1.32	0.032	0.740
sig	0.371	0.265	0.847	0.253	0.984	0.367	0.502	0.224	0.756	0.280	0.968	0.533
TRA F	0.461	21.0	0.390	24.6	0.652	5.16	1.69	16.76	1.91	18.0	0.583	4.01
sig	0.500	0.00*	0.535	0.00*	0.423	0.009*	0.196	0.00*	0.161	0.00	0.562	0.013
EXP F	4.28	40.5	2.16	66.44	0.557	18.33	2.29	52.1	2.88	64.8	0.186	18.6
sig	0.044	0.00*	0.148	0.00*	0.459	0.00*	0.113	0.00*	0.067	0.00*	0.831	0.00*
INS F	0.429	2.12	2.29	3.56	3.49	1.89	0.347	1.68	1.12	2.71	1.46	1.14
sig	0.516	0.132	0.317	0.037	0.068	0.162	0.708	0.186	0.335	0.056	0.242	0.345
IMP F	3.68	7.70	1.36	0.726	1.29	2.31	1.73	5.41	1.08	0.763	1.07	1.93
sig	0.061	0.001*	0.249	0.489	0.262	0.110	0.188	0.003	0.347	0.521	0.352	0.139

F: F-statistics.
sig: significance level.

tion (4). (This is not reported here; see Takatoshi Ito, 1988b.) Exporters have a depreciation bias in their expectation formation compared to importers and trading companies for any horizon. Table 2, panel B shows that, for any horizon, about half of the 44 individuals have a significant bias in their forecasts. The deviations are sometimes very large.

One might object to a formulation of the individual effects in the form of biases in the constant term. They could have different models. Since it is not likely that the JCIF or the econometrician could persuade each forecaster to justify the forecast with a model every week, we have to guess the form, assuming that each market participant has a common autoregressive forecasting model, but with different coefficients on the lag terms (possibly because of differences in their prior beliefs). As discussed above, idiosyncratic coefficients can be estimated from equation (6). The results are shown in Table 3.

Table 3 once again shows that exporters and trading companies are significantly heterogeneous for each of the three horizons. However, the differences came from the biases in the individual (constant term) effects, not from the idiosyncratic coefficients of the lagged variables. Importers for the one-month horizon and banks for the six-month horizon also show the individual (constant) effect, as in Table 2, panel A, but fail to show the idiosyncratic coefficients on the lagged variables. Therefore, the heterogeneity is more like a constant bias rather than the differences in reacting to the recent changes in the exchange rate. Table 2 and Table 3 show solid evidence for heterogeneous expectation formations among market participants.

C. *Discussion*: *Heterogeneity and Rational Expectations*

Most of the modern theory of finance and macroeconomics assumes the existence of a representative agent whose decision is an aggregate of market participants. In fact, the hypothesis of rational expectations would require that market participants be homogeneous in their formation of expectations since the true stochastic process is unique. Therefore, findings of heterogeneity in this section cast some doubts on the homogeneous agent framework commonly used in finance and macroeconomics.

One might argue that if agents have private information that econometricians do not observe, the existence of individual effects may not be inconsistent with rational expectations. However, important news and variables in the foreign exchange market are generally common knowledge. In fact, even if the individual information sets are different, the difference in expectations conditional on a common (i.e., intersection of) information set should be unbiased. The constant term, which detects individual effects, is certainly contained in the common information set. Thus, the finding of significant individual biases rejects rational expectations.[5]

Put differently, under the assumption of rational expectation but private information, the forecast differences across individuals, that is, the dependent variables in equations (3) and (4), must be serially uncorrelated, contrary to our findings, provided that lagged group average forecasts are part of the common information set (which is the case in the JCIF survey as explained in Section II, Part A.)

One possible explanation of heterogeneity consistent with rational expectations would be a slow learning process due to a strongly biased prior. However, one has to model a learning process to assert this. Then, we would be able to discuss how biases can be related to individual priors and learning processes. This is beyond the scope of this paper.

D. *Discussion*: *Wishful Expectations*

Having established heterogeneity, a discussion of why certain market participants have depreciation or appreciation biases is in order. From Table 2, we notice some regularity in the group effects: in the one-month-ahead forecasts, exporters have a depreciation bias, while importers have an appreciation bias. The exporters' forecasts show a continuing deviation from the mean, significantly biased toward a yen depreciation, as the forecast horizon lengthens. In the three-month- and six-month-ahead predictions, trading companies, as opposed to importers, show a bias toward appreciation.

Exporters tend to be long in dollars and importers short in dollars. It is difficult to completely cover the exposure to the foreign exchange risk, since the forward markets exist only up to a one-year horizon, and timings of trade and financial transactions cannot be matched exactly.

Therefore, exporters wish that the yen will depreciate in the future, enabling their profit margins to increase and their products to compete better in the foreign markets. (This argument rests on an assumption of incomplete "pass-through," which is documented, for example, by Paul Krugmen, 1987, and Kenichi Ohno, 1988.) Their responses, being biased toward a yen depreciation relative to the average, seem to agree with their wishes.

On the contrary, importers' responses reflect their wish for a stronger yen so that their import costs will decrease given incomplete path-through. Note that the group effect of trading companies behaves like that of import industries. One might think that the change in the exchange rate may be neutral for trading companies, since they are just intermediaries of imports and exports. However, the leading Japanese trading companies handle more imports than exports. In 1983, the revenues of the leading nine trading companies were derived from export-oriented activities for 20.0 percent, import-oriented activities for 23.6 percent, domestic activities for 40.3 percent, and trade between foreign countries for 16.1 percent (Miyohei Shinohara, 1986, p. 164).

Hence, the findings show that market participants apparently form "wishful expectations." (A "Chicago test" for the validity of survey data would be to check whether money is where the mouth is. But the result here shows that people "put the mouth where the money is.") This "wishful expectation"

[5] I owe the observation in the paragraph to an anonymous referee. There have been some investigations examining whether diverse expectations can be rational depending upon agents' information sets (see, for example, Feldman, 1987; Marcet and Sargent, 1989; and Frydman, 1982, 1987).

—or an "optimist" view in John D. Hey's (1984) sense—may be a reflection of nonrational honest mistakes in expectation formation. A straightforward interpretation would be for respondents to mix wishful thinking with objective forecasts.[6] However, there are a few deeper explanations of wishful expectations.[7]

The Japanese manufacturing and trading companies usually set an in-house exchange rate for internal accounting, and the rate can be used for coordinating the sales department with the other departments. It is possible that these in-house rates are heterogeneous, and moreover are slightly biased so that the sales department is encouraged. The survey responses from these companies may be influenced by the biased in-house exchange rate, although the respondent is not from the sales department.

If the announcement of the JCIF survey is very influential on the market, the respondent may be induced to try manipulating the announced survey result by answering with biased forecasts. Exporters respond to the JCIF by announcing a depreciated rate, but only slightly depreciated so as to avoid obvious detection, in the hope that the survey mean is biased toward depreciation. Exporters hope that the mean expectation with an "unexpected" depreciating bias could cause others to start selling yen, thus creating a self-fulfilling prophecy; if importers understand that exporters have incentives to lie, then importers would counter by manipulating their announcements; and vice versa. Thus, as a Nash equilibrium, the mean may not be biased after all, although exporters and importers are biased.

Despite its appeal to economists who are trained to think seriously about expectation and manipulation, this story of a manipulative motive has a few shortcomings. First, the size of survey, that is, 44 respondents, is large enough that a manipulation by one respondent is insignificant unless the bias is large enough to be easily detected by the JCIF. Second, if other participants understand that exporters and importers have incentives to lie, then they would not take the JCIF survey seriously, thereby removing the incentive to lie. It may be the case that market participants are simply naive in forming wishful expectations.

III. Rationality of Expectations

A. *Tests of Unbiasedness and Orthogonality*

In this section, I will apply standard tests of rational expectations to this survey data.[8] First, if the forecasts are rational, the fore-

[6] One might think that intelligent people like professional traders and dealers can separate wishful thinking from scientific forecasts. However, there is some evidence in the psychology literature, kindly suggested by Kenneth J. Arrow, that wishful thinking is rather common in social cognition and views of the self.

> Theories of the causal attribution process, prediction, judgments of covariation, and other tasks of social inference incorporated the assumptions of the naive scientists as normative guidelines with which actual behavior could be compared.
>
> It rapidly became evident, however, that the social perceiver's actual inferential work and decision making looked little like these normative models. Rather, information processing is full of incomplete data gathering, shortcuts, errors, and biases. In particular, *prior expectations and self-serving interpretations* weigh heavily into the social judgment process. (Taylor and Brown, 1988, p. 194, with emphasis added)

[7] One might think that there may be self-selection among entrepreneurs and dealers: Those who are optimistic about the yen appreciation (depreciation) develop import (export, resp.) business. However, the JCIF polls include only leading companies, so that it is difficult to imagine that they change their types of business because of exchange rate expectations. Those who are in charge of foreign exchange expectations and trades in those companies are usually in-house staff, who are subject to a lifetime employment practice. It is hardly the case in Japan that foreign exchange professionals hop companies according to their biases in expectations.

[8] For the aspects of econometrics, see Mishkin (1983). The same procedure has been applied to the MMS data by Dominguez (1986). In this paper, I assume that reported forecasts in the survey are the subjective means of respondents. However, if agents were reporting the medians of a skewed subjective distribution, then the results of rationality tests could be affected.

cast errors should be random. In other words, survey forecasts should be unbiased. Second, given rational expectations, forecast errors should be uncorrelated with (orthogonal to) any information available at the time the forecast is made. Otherwise, the variable correlated with the *ex post* error could have been exploited to make a better forecast.

Under the null hypothesis of rational expectations, the realized spot rate is the sum of a forecast and a forecast error:

$$(7)\quad s(t+k)=s^e(t,k)-h(t,k),$$

where $h(t,k)$ is the mean zero forecast error, uncorrelated with any variables available at t. It is well known that forecast errors, $h(t,k)$, would be serially correlated if the forecast horizon is longer than the observational frequency, that is, $k>2$. Therefore, rational expectations imply that $a=0$ and $b=1$ in the following regression:

$$(8)\quad s(t+k)-s(t) = a+b(s^e(t,k)-s(t))+u(t).$$

The test statistics are calculated using the Generalized Method of Moments to take care of the serial correlations of $u(t)$. Results of this unbiasedness test are reported in Table 4, panel A.

Unbiasedness is rejected for trading companies and insurance companies of the one-month horizon, for securities and import companies of the three-month horizon, and for all groups but banks and import industries for the six-month horizon. These rejections are evidenced for rejecting a rational expectation hypothesis, in that market participants had unbiased forecasts. We would miss some rejections if we were only to look at the average of the 44 participants, since for the one-month and three-month horizons, rejections by some groups are not detected in the average for all participants.[9]

The second implication of rational expectations is the orthogonality: Under the null hypothesis, forecast errors, $h(t,k)=s^e(t,k)-s(t+k)$, are uncorrelated with any information, $z(t)$, at time t. In the literature, the past forecast errors $s^e(t-k,k)-s(t)$; the forward premium, $f(t,k)-s(t)$; or the recent actual change $s(t-k)-s(t)$ have been popular candidates for variables in the information set. I will follow the standard procedure by regressing the *ex post* forecast errors on these candidate variables:

$$(9)\quad s^e(t,k)+s(t+k) = a+b(z(t)-s(t))+e(t),$$

where $z(t)=s^e(t-k,k)$, $f(t,k)$, $s(t-k)$. Rational expectations (orthogonality) is a null hypothesis of $a=b=0$. Results of the estimation of equation (9), with $z(t)=s(t-k)$, and the test of null hypothesis is reported in Table 4, panel B. (Results for other cases of $z(t)$ are essentially the same and not reported here. See Ito, 1988.) There are only a few instances of rejections of the one-month and three-month horizons. However, for the six-month horizon, the rejection is unanimous. This is consistent with the results of unbiasedness tests. So far, there is little evidence rejecting the rational expectation hypothesis for the shorter horizons.

Variables in the information set are not restricted to those tested above. When the second lagged term is added, the number of rejection cases increases dramatically. The results of estimating the following equation are reported in Table 4, panel C:

$$(10)\quad s^e(t,k)-s(t+k) = a+b_1(s(t-k)-s(t)) + b_2(s(t-k-1)-s(t-1)) + e(t).$$

Table 4, panel C shows rejections for most groups in all horizons. Even if the orthogonality test is conducted at the individual level, about three-quarters of the individuals are judged to be irrational.

[9] However, Muth (1961) originally interpreted rational expectations as applying only to aggregate expectations.

TABLE 4—TESTS OF RATIONAL EXPECTATIONS

(A) Unbiasedness

(i) Estimates and Standard Errors of *a* and *b*; Chisq and signif. for *AVE*

1 Month			3 Month			6 Month		
a	*b*	*CHISQ*	*a*	*b*	*CHISQ*	*a*	*b*	*CHISQ*
−0.028 (0.017)	−0.485 (0.969)	2.59 0.274	−0.043 (0.034)	1.167 (1.167)	5.21 0.074	−0.119 (0.041)	0.908 (0.741)	10.09 0.006*

(ii) Number of Cases in Group Data

	1 Month	3 Month	6 Month
Fail to reject *H* (at 1 percent)	4	4	2
Reject *H* (at 1 percent)	2 (TRA, INS)	2 (SEC, IMP)	4 (SEC, TRA EXP, INS)

(B) Orthogonality, past exchange rate movement, with 1 lag

(i) Estimates and Standard Errors of *a* and *b*; Chisq and signif. for *AVE*

1 Month			3 Month			6 Month		
a	*b*	*CHISQ*	*a*	*b*	*CHISQ*	*a*	*b*	*CHISQ*
0.004 (0.010)	0.166 (0.203)	3.883 0.144	0.042 (0.025)	0.306 (0.225)	9.504 0.009	0.114 (0.036)	0.227 (0.358)	18.908 0.000

(ii) Number of Cases in Group Data and Micro Data

	Group Data			Micro Data		
	1 Month	3 Month	6 Month	1 Month	3 Month	6 Month
Fail to reject *H*	5	3	0	37	26	11
Reject *H* (at 1 percent)	1 (EXP)	3 (SEC, EXP, IMP)	6	7	18	33

(C) Orthogonality, past exchange rate movement, with 2 lags

(i) Estimates and Standard Errors of *a* and *b*; Chisq and signif. for *AVE*

1 Month				3 Month				6 Month			
a	*b*	*c*	*CHISQ*	*a*	*b*	*c*	*CHISQ*	*a*	*b*	*c*	*CHISQ*
0.007 (0.011)	0.247 (0.185)	−0.323 (0.207)	38.29 (0.000)	0.043 (0.025)	0.330 (0.220)	−0.095 (0.093)	14.35 (0.002)	0.112 (0.034)	0.183 (0.299)	0.174 (0.342)	21.75 (0.000)

(ii) Number of Cases in Group Data and Micro Data

	Group Data			Micro Data		
	1 Month	3 Month	6 Month	1 Month	3 Month	6 Month
Fail to reject *H*	0	2 (BAN, TRA)	1 (BAN)	11	22	21
Reject *H* (at 1 percent)	6	4	5	33	22	23

B. *Discussion: Peso Problem and Bubbles*

Failing the rationality test in small samples may not imply that expectations are formed irrationally, due to the often involved caveats of peso problems and bubbles.[10] (See Maurice Obstfeld, 1987, and George W. Evans, 1986, and references thereof for discussions of these issues.) Suppose that conditional forecasts were formed rationally taking into account a small probability of "crash," but that the crash did not occur in the (small) sample. Then forecasts appear to have been biased when judged from *ex post* forecast errors. This is known as the peso problem. The sample size of this study is admittedly small (about two years), and this could be a reason for a rejection of rationality.

However, the sample period for this study includes a volatile period after the Plaza

[10] If the forward rate is used in place of expectation of survey data, as is the case in papers other than ones with survey data, risk aversion is another source of bias in forecast errors

Agreement of September 1985. (See Takatoshi Ito, 1987, for news analysis of the exchange rate volatility after the Plaza Agreement.) The process of the sharp yen appreciation after the Plaza agreement can be regarded as a long-awaited "crash" of the dollar value. However, market expectations underestimated the magnitude of this crash.[11]

In summary, this section shows that most of the market participants violate necessary conditions of the rational expectations hypothesis. However, these results should be interpreted with caution, because they could be a case of a peso problem.

IV. Expectational Twist

A. *Introduction to Twist and Consistency*

In this section, the consistency of expectation formation of short- vs. long-term expectations, as discussed in Froot and Ito (1989), is explored. Frankel and Froot (1987b) showed that the short-term expectations are of the bandwagon type, while the long-term expectations show some regressive characteristics.[12] Thus, I will first replicate their regressions, and then raise the question of how to interpret a "twist" found in the data.

However, Frankel and Froot (1987b) ignored the consistency issue of short- and long-term expectations formation: If expectations formation is internally consistent, a long-term forecast should be identical to the results of sequential substitutions of short-term forecasts, given a function of expectations formation. The consistency becomes a testable hypothesis in the form of cross-equation constraints on the coefficients of the short- and long-term forecast equations. This consistency problem is parallel to the cross-equation constraints implied in the context of the interest rate term structure (Thomas Sargent, 1979) and in the context of uncovered interest parity (Ito, 1988a; Takatoshi Ito and Danny Quah, 1989). Froot and Ito (1989) have applied the test of consistency to the data collected by Money Market Service (MMS) for one-week- and one-month-ahead forecasts and the Economist Financial Report for three-, six-, and twelve-month forecasts. They also used the averages from the JCIF data. In this paper, the same test is applied to the group means of the JCIF data, where one-, three-, and six-month forecasts are available.

B. *An Example of Extrapolative Expectation with One Lag*

First, let us consider, following Frankel and Froot (1987b), the extrapolative expectation with one lag:

$$(11)\quad s^e(t,k)-s(t) = a+b(s(t-1)-s(t))+e(t).$$

In (11), $b<0$ implies a (destabilizing) bandwagon effect, while $b>0$ implies a stabilizing expectation formation. Results are reported in Table 5, which shows that the 1 percent yen appreciation would make the average individual expect a further 0.01 percent appreciation in one month. However, the table also implies that the shock would make the same individual form an expectation of a 0.13 percent depreciation in three months and a 0.22 percent depreciation in six months. Although different groups have different biases, the pattern of coefficients,

$$b(\text{one month}) < b(\text{three months}) < b(\text{six months}),$$

[11] In that sense it may seem inappropriate to invoke the peso problem explanation in the usual sense for this period. The biased forecast errors resulting from the underestimation of the magnitude of a crash could be called the "Plaza problem." Both "peso problems," which arise when an infrequent crash did not happen, and "Plaza problems," which occur when an infrequent crash did happen, are small sample problems. Moreover, the latter is a special case of peso problems: A policy switch, including interventions, could halt a dollar decline and reverse the movement with a small probability. That did not happen in the small sample.

[12] Frankel and Froot (1987b) showed, using the MMS, the Economist, and the AMEX data sets, that short- and long-term expectations seem to have different characteristics. The data set with the short-term horizon yields the estimates indicating a bandwagon type (extrapolative) effect, while the data set with the long-term horizon yields results with a more regressive nature. However, the direct comparison of the short-term and long-term horizons is limited in their study because of the spread of horizons across different data sets and different sample periods.

TABLE 5—EXPECTATION FORMATION, EXTRAPOLATIVE EXPECTATION WITH ONE LAG
$s^e_j(t,k)-s(t)=a+b\{s(t-1)-s(t)\}+E(t)$

Cases: $b<0$ belief in a bandwagon effect
$b=0$ belief in constant appreciation
$b>0$ distributed lag form
H: $a=b=0$ belief in random walk

Estimates of a and b and their (standard errors)
CHISQ for Hypothesis *H*: *chisq*$(df=2)$ and (significance level)
AR1 process on *E* is assumed. *RHO* is not reported here.

	Horizon (k)								
		1 Month			3 Month			6 Month	
	a	b	*CHISQ*	a	b	*CHISQ*	a	b	*CHISQ*
AVE	−0.015	−0.011	49.42	−0.017	0.137	9.60	−0.002	0.220	5.49
	(0.002)	(0.035)	0.000	(0.005)	(0.050)	0.000	(0.009)	(0.066)	0.002
BAN	−0.014	−0.008	62.85	−0.019	0.087	6.75	−0.011	0.134	2.06
	(0.001)	(0.044)	0.000	(0.005)	(0.056)	0.003	(0.009)	(0.077)	0.139
SEC	−0.011	−0.058	8.05	−0.011	0.149	2.46	0.006	0.224	1.69
	(0.003)	(0.061)	0.001	(0.005)	(0.108)	0.097	(0.009)	(0.141)	0.195
TRA	−0.020	−0.029	69.67	−0.027	0.067	21.42	−0.011	0.194	2.41
	(0.002)	(0.068)	0.000	(0.004)	(0.096)	0.000	(0.006)	(0.120)	0.101
EXP	−0.009	0.061	18.82	−0.004	0.168	3.19	0.016	0.304	6.77
	(0.001)	(0.039)	0.000	(0.004)	(0.068)	0.050	(0.010)	(0.095)	0.003
INS	−0.018	0.015	17.20	−0.027	0.237	18.42	−0.001	0.376	5.79
	(0.003)	(0.067)	0.000	(0.005)	(0.068)	0.000	(0.010)	(0.111)	0.006
IMP	−0.018	−0.134	28.43	−0.019	0.285	6.95	−0.008	0.288	3.60
	(0.003)	(0.075)	0.000	(0.006)	(0.108)	0.002	(0.008)	(0.110)	0.035

Number of Cases in Micro Data Horizon	1 Month	3 Month	6 Month
$b \gg 0$ sig.	3	8	10
$b > 0$ insig.	18	29	31
$b < 0$ insig.	20	7	3
$b \ll 0$ sig.	3	0	0

is almost unanimously observed. Hence, we may draw a conclusion, similar to that of Frankel and Froot (1987b), that the long-term expectation is more stabilizing than the short-term expectation.

It is easy to show that so long as the extrapolative expectation with one lag is assumed, a twist, that is, an appreciation in the short run and a depreciation in the long run, in expectation is impossible. Put differently, the assumed formulation is not rich enough for the observed twist to be consistent.

C. *Consistency Tests*

Next, we adopt a distributed lag expectation formulation with more than two lags, a formulation rich enough to produce a twist in expectation. Consider estimating the following k-month $(k=1,3,6)$ expectation formations:

$$s^e(t,k)=d_k+(1+a_k)s(t)+b_k s(t-1)+c_k s(t-2)+u_k(t), \tag{12}$$

where $u_k(t)$ are independent, random variables representing observation errors. After substitution, using the iterated projection (see Froot and Ito, 1989), the consistency restrictions as cross-equation constraints are derived (see Table 6).

Each of two sets of cross-equation restrictions, one-month vs. three-month, and three-month vs. six-month, is tested separately, and the results are reported in Table

TABLE 6—CONSISTENCY TESTS, ONE-MONTH VS. THREE-MONTH EXPECTATIONS

$s_j^e(t,1) - s(t) = d_1 + a_1 s(t) + b_1 s(t-1) + c_1 s(t-2)$

$s_j^e(t,3) - s(t) = d_3 + a_3 s(t) + b_3 s(t-1) + c_3 s(t-2)$

H: Consistency restrictions:

$$\begin{aligned} d_3 &= (2 + a_1 + b_1 + ((1+a_1)^2))d_1 \\ a_3 &= c_1 - 1 + (2(1+a_1)b_1) + ((1+a_1)^3) \\ b_3 &= ((1+a_1)c_1) + (b_1^2) + (b_1((1+a_1)^2)) \\ c_3 &= (c_1((1+a_1)^2)) + (b_1 c_1) \end{aligned}$$

			Estimates and (Standard Errors)						
		1 Month (*OLS*)				3 Month (*GMM*)			*H*
	d_1	a_1	b_1	c_1	d_3	a_3	b_3	c_3	*CHISQ*
AVE	−0.0261 (0.0050)	0.0003 (0.0001)	0.0001 (0.0002)	−0.0003 (0.0001)	−0.0254 (0.0071)	−0.0008 (0.0071)	0.0005 (0.0001)	0.0003 (0.0001)	2182.1 0.000

In the Group data, H is rejected in all 6 groups.
In Micro data, H is rejected in 42 out of 44 individuals.

Three-Month vs. Twelve-Month Expectations

$s_j^e(t,3) - s(t) = d_3 + a_3 s(t) + b_3 s(t-3) + c_3 s(t-6)$

$s_j^e(t,6) - s(t) = d_6 + a_6 s(t) + b_6 s(t-3) + c_6 s(t-6)$

H: Consistency Restrictions:

$$\begin{aligned} d_6 &= (2 + a_3)d_3 \\ a_6 &= (1+a_3)^2 + b_3 - 1 \\ b_6 &= (1+a_3)b_3 + c_3 \\ c_6 &= (1+a_3)c_3 \end{aligned}$$

		3 Month (*OLS*)				6 Month (*GMM*)			*H*
	d	*a*	*b*	*c*	*D*	*A*	*B*	*C*	*CHISQ*
AVE	0.0218 (0.0220)	−0.0009 (0.0003)	0.0004 (0.0003)	0.0002 (0.0002)	0.0508 (0.0285)	−0.0019 (0.0003)	0.0006 (0.0002)	0.0008 (0.0001)	570.1 0.000

In Group data, H is rejected for all 6 groups.
In Micro data, H is rejected for 42 out of 44 individuals.

6. The consistency is overwhelmingly rejected in this formulation, too.

D. *Discussion*: *Inconsistency*

I hasten to add a caveat. If we misspecify the expectation formation, then the results in this section are not valid. For example, if a policy switch, such as a monetary tightening, is expected to occur around the second month from the point of forecasting, it is "consistent" to have a twist, although the test in this paper would not capture it.

One might think that people use different economic variables for forecasting the future spot rate with different horizons. For example, chart (technical) analysis, which is a special case of (univariate) distributed lag expectation formations, is used for the short-term horizon, but other factors come into consideration for the long-term horizon. A list of other factors includes trade balances, inflation rate differentials, interest rate differentials, fiscal deficits, and policy switches. However, if these factors are relevant in the long run, they should be relevant in the short run, although the effect may be small in the short-run.[13]

[13] Suppose that uncovered interest parity (no risk premium) holds. An interest rate differential of 6 percent implies that the exchange rate changes by approximately 3 percent in six months, a significant and easily detectable change. However, it predicts only a 0.5 percent change in one month, a change that is small and may escape detection.

V. Concluding Remarks

In this paper, newly available survey data on the expected exchange rate in the Tokyo market were used to test several hypotheses regarding expectation formations. The JCIF data set is better than the data sets previously used by Frankel and Froot (1987a,b) in that the survey includes the expectations of different industries, not only of banks and financial institutions but also of exporters and importers. Moreover, individual responses can be used to avoid the aggregation problem altogether.

Following are the major findings of this paper: First, market participants are heterogeneous, with constant-term biases in their expectation formations. Second, "wishful expectations" were found: exporters (importers) are biased toward yen depreciation (appreciation) relative to others. Third, when the usual rationality tests were applied, among different groups, the unbiasedness of expectation was rejected in a few instances for shorter horizons and unanimously rejected in the six-month horizon. Orthogonality was soundly rejected. We may conclude that we have strong evidence against rational expectation formation in the Tokyo foreign exchange market. Fourth, consistency is overwhelmingly rejected given that the expectation formation is a distributed lag structure with two lags.

The present paper suggests that it is important to consider a model with heterogeneous agents for the international financial market. I hope that this paper stimulates the research in this direction.

REFERENCES

Dominguez, Kathryn M., "Are Foreign Exchange Forecasts Rational?" *Economics Letters*, 1986, *21*, 277–81.

Evans, George W., "A Test for Speculative Bubbles in the Sterling-Dollar Exchange Rate: 1981–84" *American Economic Review*, September 1986, *76*, 621–36.

Feldman, Mark, "An Example of Convergence to Rational Expectations with Heterogeneous Beliefs," *International Economic Review*, October 1987, *28*, 635–50.

Figlewski, Stephen and Paul Wachtel, "The Formation of Inflationary Expectation," *Review of Economics and Statistics*, February 1981, *63*, 1–10.

Frankel, Jeffrey A. and Froot, Kenneth A., (1987a) "Short-Term and Long-Term Expectations of the Yen/Dollar Exchange Rate: Evidence from Survey Data," *Journal of the Japanese and International Economies*, September 1987, *1*, 249–74.

______, (1988b) "Foreign Exchange Rate Expectations: Micro Survey Data," National Bureau of Economic Research, Working Paper, no. 2679, August 1988.

Froot, Kenneth A. and Ito, Takatoshi, "On the Consistency of Short-Run and Long-Run Exchange Rate Expectations," *Journal of International Money and Finance*, December 1989, *8*, 487–510.

Frydman, Roman, "Towards an Understanding of Market Processes: Individual Expectations, Learning, and Convergence to Rational Expectations Equilibrium," *American Economic Review*, September 1982, *72*, 652–68.

______, "Diversity of Information, Least Squares Learning Rules and Market Behavior," mimeo., New York University, Department of Economics, 1987.

Hey, John D., "The Economics of Optimism and Pessimism: A Definition and Some Applications" *Kyklos*, 1984, *37*, 181–205.

Ito, Takatosi, "The Intradaily Exchange Rate After the Group of Five Agreement," *Journal of the Japanese and International Economies*, September 1987, *1*, 275–98.

______, (1988a) "Use of (Time-Domain) Vector Autoregressions to Test Uncovered Interest Parity," *Review of Economics and Statistics*, May 1988, *70*, 296–305.

______, (1988b) "Foreign Exchange Rate Expectations: Micro Survey Data," National Bureau of Economic Research, Working Paper no. 2679, August 1988.

______ **and Danny Quah**, "Hypothesis Testing with Restricted Spectral Density Matrices, with an Application to Uncovered Interest Parity," *International Economic Review*, February 1989, *30*, 203–15.

Krugman, Paul, "Pricing to Market When the Exchange Rate Changes," in S. W. Arndt and J. D. Richardson eds., *Real-Financial Linkages Among Open Economies*, Cambridge, MA: MIT Press, 1987.

Marcet, Albert and Sargent, Thomas J., "Convergence of Least Squares Learning in Environments with Hidden State Variables and Private Information," *Journal of Political Economy*, December 1989, *97*, 1306–22.

Mishkin, Frederic S., *A Rational Expectations Approach to Macroeconometrics*, Chicago: University of Chicago Press, 1983.

Muth, John F., "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

Obstfeld, Maurice, "Peso Problems, Bubbles, and Risk in the Empirical Assessment of Exchange-Rate Behavior," National Bureau of Economic Research, Working Paper, no. 2203, April 1987.

Ohno, Kenichi, "Export Pricing Behavior of Manufacturing: A U.S.–Japan Comparison," International Monetary Fund, mimeo., June 1988.

Sargent, Thomas, "A Note on Maximum Likelihood Estimation of the Rational Expectations Model of the Term Structure," *Journal of Monetary Economics*, 1979, *5*, 133–43.

Shinohara, Miyohei, ed., *Lectures on the Japanese Economy* (in Japanese), Tokyo: Keizai Shinpo Sha, 1986.

Taylor, Shelley E. and Brown, Jonathon D., "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," *Psychological Bulletin*, 1988, *103*, 193–210.

A Non-Parametric Analysis of Productivity: The Case of U.S. and Japanese Manufacturing

By JEAN-PAUL CHAVAS AND THOMAS L. COX*

A non-parametric analysis of technology, technical change, and productivity is presented in the context of cost minimizing behavior. A number of non-parametric tests concerning the existence and nature of technical change are applied to U.S. and Japanese manufacturing data. Non-parametric measures of technical change are presented. (JEL 226)

The issues of productivity measurement and technical change have received considerable attention in the literature (for example, Dale W. Jorgenson and Zvi Griliches, 1967; Ryuzo Sato, 1970; Hans Peter Binswanger, 1974; Rodney E. Stevenson, 1980; John R. Norsworthy and David H. Malmquist, 1983). Using a primal approach, technical progress is typically measured in terms of changes in output that are not attributable to changes in inputs. Alternatively, dual approaches measure technical progress in terms of the contributions to changes in cost (profit) not attributable to changes in input prices and output levels (output prices). This suggests a need to distinguish between the contributions of technical progress and those of returns to scale (for example, George J. Stigler, 1961) and input prices (for example, Zvi Griliches, 1958).

However, in the absence of a priori hypotheses concerning the structure of technical change, Peter Diamond et al. (1978) have shown the non-identifiability of the elasticities of substitution and the bias in technical change.[1] This suggests that a set of production data can be generated by *more than one combination* of technology and technical change and raises serious identification and measurement issues. Depending on the data as well as maintained hypotheses concerning the structure of technology, the measurement of technical change may be exactly identified, identified up to a range of indeterminacy, or not identified at all (Diamond et al.). This identification problem suggests that traditional parametric analysis of technology and technical change may give results that are sensitive to the particular parametric specification utilized.[2] In this context, analyses of technology change that are less dependent on the parametric specification of the model could be insightful.

The objective of this paper is to present a non-parametric analysis of technology, technical change and productivity in the context of cost minimizing behavior. The paper first extends the non-parametric analysis of production decisions (Sydney N. Afriat, 1972; Giora Hanoch and Michael Rothschild, 1972; Hal R. Varian, 1984) in the context of technical change using an augmentation hy-

*Respectively Professor and Assistant Professor of Agricultural Economics, University of Wisconsin-Madison, WI 53706. The seniority of authorship is equally shared. We wish to thank two anonymous reviewers for very helpful comments. Remaining errors are the joint responsibility of the authors.

[1] Technical change is said to be Hicks neutral if the marginal rate of substitution between inputs is not affected by the change (for example, see Charles Blackorby et al. 1976). Non-neutral technical change is said to be biased. For example, biased technical change is labor saving (capital using) if the marginal product of capital rises relative to the marginal product of labor, *ceteris paribus*.

[2] Note that, in general, the use of flexible functional forms (for example, translog or generalized Leontief for a production, cost, or profit function) does not help solve the identification of technical change problem. What is needed to solve the identification problem is a priori information about technology or the nature of technical change (see Diamond et al.).

pothesis, that is, where technical progress is assumed to increase the effectiveness of inputs in the production of output (Section I). In particular, assuming cost minimization, a series of non-parametric tests for the existence and nature of technical change are derived (Section II). Empirical implementation of these non-parametric tests is discussed in Section III. An application of the proposed methodology to U.S. and Japanese manufacturing data from Norsworthy and Malmquist is presented in Section IV.

This analysis provides useful insights on the nature of technology and technical change. First, it allows an empirical investigation of various separability hypotheses concerning the production function. Second, non-parametric tests of Hicks neutral technical change are presented. Third, non-parametric measures of productivity changes based on the augmentation hypothesis are obtained. Although our approach does not solve the identification problem raised by Diamond et al., our results do not depend on a particular parametric specification of production relationships. Such results are contrasted with the corresponding parametric results obtained by Norsworthy and Malmquist using the same data. Finally, under the augmentation hypothesis, non-parametric measures of technical change can be calculated. The measures obtained for U.S. and Japanese manufacturing appear reasonable and illustrate the empirical usefulness of the proposed methodology.

I. Non-Parametric Tests of Production Decisions

In order to develop non-parametric tests of production, consider a cost-minimizing firm facing the following problem:

$$(1)\quad C(p,y,A)=\min_x\{p'x: f(y,A) \le g(x,A),\ x\ge 0\},$$

where x is an $(n\times 1)$ input vector, p is the $(n\times 1)$ vector of corresponding prices, y denotes output, and $f(y,A)\le g(x,A)$ represents technology, $A>0$ being a vector of technology indices. The function $g(x,A)$ is assumed to be strictly increasing and concave in x throughout the paper. Also, we assume the existence of a point $\bar{x}\ge 0$ satisfying $f(y,A)<g(\bar{x},A)$ (M. Slater's 1951 condition). Equation (1) characterizes the production decisions of a firm that behaves in a way consistent with the cost minimization hypothesis, $C(p,y,A)$ being the indirect cost function.

Assume that the firm is observed choosing x T times, $x_1,\dots,x_T$, each observation being a feasible point (i.e., satisfying $f(y_t,A_t)\le g(x_t,A_t)$, $x_t\ge 0$) associated with a situation t characterized by input prices p_t, output y_t and technology A_t, $t=1,\dots,T$. It is of interest here to investigate under what conditions the decision set $\Omega=\{x_1,\dots,x_T\}$ is consistent with cost minimization as stated in (1). Throughout the paper, we take cost minimization as a maintained hypothesis. In this context, we present non-parametric tests of production decisions. These non-parametric tests consist in checking the consistency of the actual decisions $\Omega=\{x_1,\dots,x_T\}$ with the cost minimization problem (1) without a parametric specification of the production technology.

A basis for these non-parametric tests is presented in the following proposition (see the proof in the Appendix).

PROPOSITION 1: *Given a set of feasible decisions* $\Omega=\{x_1,\dots,x_T\}$, *each* x_t *corresponding with a situation* (p_t,A_t,y_t), $t=1,\dots,T$, *then*:

a) *If* x_t *solves* $\min\{p_t'x: f(y_t,A_t)\le g(x,A_t),\ x\ge 0\}$, $g(x,A)$ *being a strictly increasing and concave function of x, then there exist* λ_t *such that*

$$(2\text{a})\quad f(y_t,A_t)-g(x_s,A_t)+\frac{p_t'}{\lambda_t}(x_s-x_t)\ge 0$$

$$(2\text{b})\quad \lambda_t>0$$

$$s,t=1,\dots,T.$$

b) *if (2a) and (2b) are satisfied, then there exists a function* $G(x,A)$ *such that* x_t *solves* $\min(p_t'x: f\{y_t,A_t)\le G(x,A_t),\ x\in \Omega\}$.

Equations (2) represent a set of necessary and sufficient conditions for the decisions $\Omega = \{x_1, \dots, x_T\}$ to be consistent with the optimization problem (1) for some production technology. Testing for consistency consists in checking whether there exists a solution to the set of inequalities in (2). Although this non-parametric test is not a statistical test (with associated probability statements), it can provide useful information on the nature of technology. More importantly, by allowing for technical change the above results extend the non-parametric analysis of production decisions proposed by Hanoch and Rothschild or Varian (1984) (see equations (5), (6), and (7) below).

The nature of technical change assumed in (1) is quite general since it does not specify a priori how the technology indexes shift the production function. However, in order to implement the non-parametric test given in (2), some knowledge of the nature of technological change (more specifically, some knowledge of the functions $f(y, A)$ and $g(x, A)$) must be hypothesized. These hypotheses must be precise enough to make (2a) empirically tractable, but without requiring a complete parametric specification of technology (otherwise we would be back in a more traditional parametric analysis of technology).

In this context, the "augmentation hypothesis" appears particularly attractive for modeling technical change: that is, technical progress increases the effectiveness of inputs in the production of output. This can be expressed by rewriting the function $g(x, A)$ as

$$g(h(x,A)) = g[h_1(x_1, A_1), h_2(x_2, A_2), \dots, h_n(x_n A_n)],$$

where $h_i(x_i, A_i)$ can be interpreted as a measure of the effectiveness of the ith input, $i = 1, \dots, n$. Defining $Y = f(y, A_0)$, the cost minimization problem (1) then becomes

$$\text{(3a)} \quad C(p, Y, A) = \min_x \{ p'x : Y \le g(h(x, A)), x \ge 0 \}.$$

We assume here that the $h_i(x_i, A_i)$ are strictly increasing functions of x_i, $i = 1, \dots, n$. Given the augmented input function $X_i = h_i(x_i, A_i)$, denote its inverse function by $x_i = H_i(X_i, A_i)$, $i = 1, \dots, n$. The above cost minimization problem can then be alternatively expressed as

$$\text{(3b)} \quad C(p, Y, A) = \min_X \{ p'H(X, A) : Y \le g(X), H(X, A) \ge 0 \},$$

where $X = (X_1, \dots, X_n)'$, $A = (A_1, \dots, A_n)'$, and $H = (H_1, \dots, H_n)'$.

Given this alternative formulation of (1) under the augmentation hypothesis, expressions (2) become

$$\text{(4a)} \quad Y_t - Y_s + (p_t'/\lambda_t) \times [H(X_s, A_t) - H(X_t, A_t)] \ge 0$$

$$\text{(4b)} \quad \lambda_t > 0$$

$$s, t = 1, \dots, T,$$

which represent necessary and sufficient conditions for production decisions to be consistent with cost-minimizing behavior. Equations (4) provide a non-parametric test of production decisions in the sense that a priori specification of the functional form $g(X)$ in the characterization of the production technology is not required. However, checking the existence of a solution to the set of inequalities in (4) requires a priori information on the functions $Y = f(y, A_0)$ and $X_i = h_i(x_i, A_i)$, $i = 1, \dots, n$.

Two simple alternative specifications for the function $f(y, A)$ and $h_i(X_i, A_i)$ are the scaling hypothesis corresponding to the multiplicative specification $Y = y/A_0$, $X_i = A_i \cdot x_i$, and the translating hypothesis corresponding to the additive specification $Y = y - A_0$, $X_i = A_i + x_i$. Assuming output translating, then $Y_t - Y_s$ in (4a) becomes $Y_t - Y_s = y_t - A_{0t} - y_s + A_{0s}$. Similarly, under input translating, we have $H(X_s, A_t) = X_s - A_t = x_s + A_s - A_t$. Substituting these expressions into (4a) yields empirically tractable non-parametric tests of production decisions. Under the maintained hypotheses of cost minimization with additive input and/or output augmenting technical change, this will provide a basis for the empirical analysis of productivity presented below.[3]

Note that the input augmentation hypotheses (either alone or with output augmentation) allow for biased technical change while output augmentation (without input augmentation) generates a Hicks neutral specification (where the marginal rate of substitution between inputs is unaffected by technical change). The next section presents non-parametric tests of production decisions and their interpretation under the additive augmentation hypothesis.

II. Non-Parametric Analysis of Technology and Productivity

Taking cost minimization as a maintained hypothesis, the above results generate non-parametric tests which can be used to investigate the nature of technology and of technical change.

A. *Weak Separability Test*

A production function is weakly separable in x if it takes the form $g[G(x), A,.]$, where $G(x)$ is an aggregator function of x, g is strictly monotonic in G, and "." represents other arguments (besides x and A) of the production function. In this case, the choice of x for a competitive firm must be made such that x_t is a solution to $\min\{p_t'x: f_t \leq G(x),\ x \geq 0\}$ for some f_t, which is a special case of (1). Then, equations (2) become

$$(5)\quad f_t - f_s + \gamma_t p_t'(x_s - x_t) \geq 0,$$

$$\lambda_t > 0 \quad s,t = 1,\dots,T,$$

where $\gamma_t = 1/\lambda_t$. This is the test for weak separability presented by Varian (1984, p. 588). If the vector x is a subset of inputs, then (5) allows non-parametric testing of the separability of selected inputs in the production function.

On the other hand, if the vector x includes all of the inputs, then (5) amounts to a test of the weak separability of all inputs from the technology indices A. But this weak separability is equivalent to stating that the marginal rate of substitution between any two inputs is independent of technical change, which is the definition of Hicks neutral technical change in its most general form (Blackorby et al.'s extended Hicks neutrality). Thus, condition (5) allows non-parametric testing of Hicks neutral technical change without an *ad hoc* specification of technology or its change.

B. *Testing for the Existence of Technical Change*

In the absence of technical change, rewrite the cost minimization problem (1) as $\mathrm{Min}\{p'x: y \leq g(x),\ x \geq 0\}$. Then, equations (2) take the form

$$(6)\quad y_t - y_s + \gamma_t p_t'(x_s - x_t) \geq 0,$$

$$\gamma_t > 0, \quad s,t = 1,\dots,T,$$

where $\gamma_t = 1/\lambda_t$. This is Varian's strong axiom of cost minimization (see Varian, 1984, p. 583). While the strong axiom of cost minimization (6) holds for any concave production function, it takes a more restrictive form under the additional assumption of constant return to scale (CRTS) technology since the marginal cost of output and average cost are constant and equal to each other. Noting that the multiplier λ in (6) is the marginal cost of output and average cost is $p'x/y$, it

[3] Alternatively, the scaling hypothesis corresponding to multiplicative input/output augmenting technical change could be used. Under the scaling hypothesis output scaling implies $Y = y/A_0$ with $Y_t - Y_s = y_t/A_{0t} - y_s/A_{0s}$, while input scaling implies $(X_i = A_i \cdot x_i)$ with $H_i(X_i, A_j) = X_i/A_j = x_i \cdot A_i/A_j$. Substituting these expressions into (4) yields non-parametric tests of production decisions. However, in the presence of input augmentations and scaling, the inequalities (4) are nonlinear in the technology indices A. The nonlinearities make the scaling hypothesis less empirically convenient than the translating hypothesis in the analysis of technical change (see Section III below). As a result, most of the results presented in this paper focus on the translating hypothesis. Selected results from the scaling hypothesis are briefly discussed in footnotes.

TABLE 1—NON-PARAMETRIC TESTS OF COST MINIMIZATION UNDER ADDITIVE AUGMENTATION (TRANSLATING) HYPOTHESES

Translating and Cost Minimization

—Output Translating ($Y = y - A_0$):

$$y_t - A_{0t} - y_s + A_{0s} + \gamma_t p_t'(x_s - x_t) \geq 0,\ \lambda_t > 0, \quad \text{where } \gamma_t = 1/\lambda_t \tag{8a}$$

—Input Translating ($X_i = x_i + A_i$):

$$\lambda_t(y_t - y_s) + \sum_{i=1}^{n} p_{it}(x_{is} + A_{is} - x_{it} - A_{it}) \geq 0,\ \lambda_t > 0 \tag{8b}$$

—Input and Output Translating:

$$\lambda_t(y_t - A_{0t} - y_s + A_{0s}) + \sum_{i=1}^{n} p_{it}(x_{is} + A_{is} - x_{it} - A_{it}) \geq 0,\ \lambda_t > 0 \tag{8c}$$

Translating and Cost Minimization Under Constant Return to Scale

—Output Translating ($Y = y - A_0$):

$$y_t \cdot \frac{p_t'x_s}{p_t'x_t} - A_{0t} - y_s + A_{0s} \geq 0 \tag{9a}$$

—Input Translating ($X_i = x_i + A_i$):

$$\sum_{i=1}^{n} p_{it}(x_{is} + A_{is} - A_{it}) - p_t'x_t \cdot \frac{y_s}{y_t} \geq 0 \tag{9b}$$

—Input and Output Translating:

$$\sum_{i=1}^{n} p_{it}(x_{is} + A_{is} - A_{it}) - \frac{p_t'x_t}{y_t}(y_s - A_{0s} + A_{0t}) \geq 0 \tag{9c}$$

then follows that $\gamma_t = 1/\lambda_t = (y_t)/(p_t'x_t)$. Hence, under CRTS (6) becomes

$$y_t - y_s + \frac{y_t}{p_t'x_t} \cdot p_t'(x_s - x_t) \geq 0. \tag{7}$$

This can be alternatively expressed as $y_t \cdot (p_t'x_s/p_t'x_t) \geq y_s$, which is Varian's non-parametric test for constant return to scale (Varian, 1984, p. 585). Thus, for a particular data set, finding that the inequalities in (7) are satisfied would imply the existence of a *stable* linear homogeneous production function that rationalizes the data in a cost minimization framework. If the inequalities in (7) are not satisfied then either the production function is not linear homogeneous, *or* the production function is not stable, *or* cost minimization is violated (or some combination of the above). In the case where linear homogeneity and cost minimization are assumed as maintained hypotheses, violations to the inequalities in (7) would provide non-parametric evidence that the production function is not stable, that is, that technological change has taken place.

C. *Testing for the Nature of Technical Change*

We have argued above that the weak separability test (5) can be used to investigate non-parametrically whether technical change is Hicks neutral or biased. However, it is often of interest to have more precise results on the nature of technical change. We now pursue this in the context of the translating hypothesis discussed in Section I. The main results are summarized in Table 1.

Under cost minimization and the translating hypothesis (where $Y = y - A_0$ for output translating and $X_i = A_i + x_i$ for input translating), inequalities (4) then yield equations (8) in Table 1. Under the additional assumption of constant return to scale, noting that the marginal cost of output is λ_t, it follows that $\lambda_t = p_t'x_t/y_t$. Cost minimization and constant return to scale then yield equations (9) in Table 1. Under cost minimization and the translating hypothesis, the inequalities (8) (or (9) under CRTS) therefore provide a basis for a non-parametric analysis of productivity and technical change. In particular, note that in the absence of input translating ($A_i = 0$, $i = 1,\ldots, n$), then output translating (the A_0's) characterizes Hicks neutral technical change.

A special case of (9c) will be of interest in this paper (see below). This involves the case of additive separability where the production function takes the form $y = g(x) + A_0 + f(Z)$, where $Z = (Z_1,\ldots, Z_m)$, $Z_i = A_i + z_i$, and x and z are inputs. This implies that the production function is strongly separable in x and z, and that technical change takes place in an output translating fashion as well as an input translating fashion but only with respect to z. Under the translating hypothesis, this is the form of the production function that is required to justify the real value added measurement of productivity (see Michael Denny and J. Douglas May, 1978). Given this form of the production function, cost-minimizing behavior implies

$$\min_x \{ w'x: \bar{g} \leq g(x),\ x \geq 0 \}$$

and

$$\min_z \{ v'z: \bar{f} \leq A_0 + f(z_1 + A_1, z_2 + A_2, \ldots),\ z \geq 0 \},$$

where w and v are the price vectors for x and z, respectively, and $y = \bar{g} + \bar{f}$. Under the assumption of constant return to scale (where $\lambda_t = (w_t'x_t + v_t'z_t)/y_t$) then, from (6) and (8c), the non-parametric results associated with this additive separability case are

$$\text{(10a)} \quad g_t - g_s + \frac{y_t}{w_t'x_t + v_t'z_t} \cdot w_t'(x_s - x_t) \geq 0$$

$$\text{(10b)} \quad \frac{w_t'x_t + v_t'z_t}{y_t} \cdot (f_t - A_{0t} - f_s + A_{0s}) + \sum_{i=1}^{m} v_{it}(z_{is} + A_{is} - z_{it} - A_{it}) \geq 0$$

$$\text{(10c)} \quad y_t = g_t + f_t.$$

Under constant return to scale, the inequalities (10) generalize Varian's test of additive separability (see Varian, 1984, p. 590) by allowing for technical change in the context of the translating hypothesis.

D. *Goodness-of-Fit Measures*

Note that each of the non-parametric tests presented above corresponds to a particular null hypothesis. If the stated inequalities are satisfied for a given set of data, then the corresponding null hypothesis would be accepted. However, if there is evidence against the null hypothesis, that is, if the stated inequalities cannot be satisfied, then it would be useful to have some measure on the strength of the evidence against the null hypothesis. Here, we propose to rely on a goodness-of-fit measure originally developed in the consumer context by Afriat (1967) and discussed in Varian (1988).

To illustrate the approach, consider the case of cost minimization under CRTS as reflected by the non-parametric test (7). Define the goodness-of-fit measures[4] $(e_1^*, \ldots, e_T^*)$ as the solutions of the following problem

$$\text{(11)} \quad \max_{e_t} \left\{ e_t: 0 \leq e_t \leq 1;\ y_t - y_s + \frac{y_t}{p_t'x_t} \cdot p_t'(x_s - e_t x_t) \geq 0,\quad s = 1,\ldots, T \right\} \quad t = 1,\ldots, T.$$

If $(1 + p_t'x_s/p_t'x_t - y_s/y_t) \geq 0$, $s = 1,\ldots, T$, note that (11) implies that $e_t^* = \text{Min}\{1, 1 + p_t'x_s/p_t'x_t - y_s/y_t\}$. If $e_t^* = 1$, $t = 1,\ldots, T$, in (11), then the inequalities in (7) are satisfied and the corresponding null hypothesis would be accepted. Alternatively if $e_t^* < 1$ for some t, then the inequalities in (7) would not be satisfied. In this case, $(1 - e_t^*)$ can be interpreted as the smallest proportional reduction in cost of production that

[4] In Afriat's terminology, e_t^* is called an efficiency index.

would have to take place before the input bundle x_t would be consistent with the non-parametric test (7). In other words, $(1-e_t^*)$ can be interpreted as the relative difference (measured in terms of relative cost of production) between the data point t and the non-parametric hypothesis (7). Thus, e_t^* in (11) can provide useful information by measuring how close the observed choices are from satisfying the null hypothesis, that is, by measuring the strength of rejection of this hypothesis. While these arguments were presented in the context of expression (7), similar results can be developed in the context of (5), (6), (8), (9), and (10). As will be illustrated in Section IV, this can provide useful goodness-of-fit measures associated with non-parametric tests.

III. Empirical Implementation

From Proposition 1, the inequalities presented in equation (5) through (10) are necessary and sufficient for the existence of a production function that would rationalize a particular set of production data under alternative hypotheses. Non-parametric testing thus involves checking the existence of a solution to these inequalities.

The empirical implementation of (7) is straightforward as the inequalities in (7) involve only observable variables, p, x, y. In this case it is a simple matter to check whether the inequalities in (7) are satisfied for all observations. However, the inequalities (5), (6), (8), (9), and (10), involve variables that are not directly observable (for example, γ_t and f_t in (5); A_{0t} and/or A_{it} in (9); A_{0t}, A_{it}, g_t, and f_t in (10)). In these cases, the non-parametric tests consist in finding whether there exists a set of values taken by the unobserved variables which would satisfy the corresponding inequalities. Since the inequalities (5), (6), (8), (9), and (10) are linear in the unobserved variables, checking the existence of a solution to these inequalities can be conveniently formulated as a linear programming problem, as discussed below.[5]

Considering the inequalities (5), (6), (8), (9), or (10), denote by q the vector of unobserved variables in these expressions. Since q shows up in linear form in these inequalities, they can be written as $B'q \geq c$, given appropriate definitions of the matrix B and the vector c. Then consider the linear programming problem

$$(12) \qquad \operatorname*{Min}_{q}\{b'q; B'q \geq c,\ q \in C\},$$

where the vector b and the feasible set C, a cone, are appropriately defined so that problem (12) is necessarily bounded. It follows that either problem (12) has a solution, or if it does not, it must be infeasible. In other words, the inequalities $B'q \geq c$ have a solution if and only if problem (12) has a feasible solution. In this context, checking the existence of a solution to the non-parametric inequalities can be done by checking the existence of a feasible solution to the linear programming problem (12) (for example, using the simplex method). By choosing appropriate values for the b's (the coefficients of the objective function in (12)), this can yield useful information concerning the rates of input and/or output augmentations that are consistent with the data (see below).

Note that, even for a moderate number of observations T, the number of constraints in the linear programming (12) will typically exceed the number of activities. In this case, it will be computationally convenient to consider the linear programming problem dual to (12)

$$(13) \qquad \operatorname*{Max}_{\bar{q}}\{c'\bar{q}: b - B\bar{q} \in C^*,\ \bar{q} \geq 0\},$$

where C^* is the polar cone of C. It is well known that (13) has an optimal solution if

[5] As mentioned in fn. 3, hypotheses involving input scaling yield non-parametric test restrictions which are nonlinear in the unobserved variables. Note that, say for $T>10$, the number of these inequalities can be large. Since solving a large system of *nonlinear* inequalities can be computationally quite difficult, this suggests that the non-parametric testing of input scaling will in general be a difficult task. In contrast, the translating augmentation hypotheses discussed here yield non-parametric tests involving *linear* inequalities which are particularly convenient for non-parametric empirical work.

and only if (12) has an optimal solution (for example, David G. Luenberger, 1984; V. A. Sposito). Alternatively, if problem (12) is infeasible, then (13) is either unbounded or infeasible.

Here, whenever the unobserved variables show up in linear form in the non-parametric inequalities, we propose to solve the dual formulation (13) (using the simplex method). If (13) has an optimal solution for particular production data, then the data are consistent with the associated hypothesis. Alternatively, if (13) is either infeasible or unbounded, then we would conclude the corresponding inequalities cannot be satisfied and that the data are not consistent with the associated hypothesis. In this case, it will be of interest to obtain the goodness-of-fit measures such as the ones proposed in (11). Again, this can be done in the context of a standard linear programming problem.

To illustrate, consider for example the case of cost minimization under input and output translating (8c). Under CRTS (where $\lambda_t = p_t'x_t/y_t$), (8c) generalizes into (9c) with goodness-of-fit measures e_t, as

$$\text{(14a)} \quad \frac{p_t'x_t}{y_t}(y_t - A_{0t} - y_s + A_{0s}) + \sum_{i=1}^{n} p_{it}(x_{is} + A_{is} - e_t x_{it} - A_{it}) \geq 0,$$

$$\text{(14b)} \qquad 0 \leq e_t \leq 1.$$

Let q be the vector of unobserved variables in (14) (i.e., the A_{0t}'s, the A_{it}'s and the e_t's). Then the set of linear inequalities in (14) can be written as $B'q \geq c$, and the solutions to these linear inequalities can be investigated in the context of the linear programming problem (12) (or (13)) given appropriate choices for the vector b. Here, we chose the b's associated with the A_i's (the input translating factors) equal to M^2, the b's associated with the A_0's (the output translating factors) equal to M, and the b's associated with the e's (the goodness-of-fit measures) equal to $-M^3$, where M is a large positive scalar. In doing so, the linear programming solution to (13) gives the largest goodness-of-fit measures $e_t^* \leq 1$ that are consistent with the data. Note that in the case where $e_t^* = 1$, $t = 1, \ldots, T$, the (14) becomes equivalent to the non-parametric test (9c). In this case, given this choice of b's, the solution to (13) gives the smallest A_i's, that is, the smallest biases in technical change, that are consistent with the data. Given the A_i's, the solution to (13) also gives the smallest A_0's, that is, the smallest productivity changes, that are consistent with the data. Thus the non-parametric approach can provide useful information in the analysis of productivity. Finally, not that while this example has focused on the empirical implementation of the non-parametric test (9c), the approach can be easily modified to handle as well the non-parametric tests discussed above.

IV. Application to U.S. and Japanese Manufacturing

In this section, we present an application of our non-parametric methodology to the analysis of productivity in U.S. and Japanese manufacturing. We apply selected results from Section II using the procedures described in Section III. The data for U.S. and Japanese manufacturing are from Norsworthy and Malmquist and include four inputs: capital, labor, energy, and material. These data cover the periods 1958–77 for the United States and 1965–78 for Japan. Implicit quantity indexes were created for output and inputs by dividing current dollar (U.S.) or yen (Japan) values by the associated price indexes. These indexes were then adjusted so that the quantity indexes are equal to 1.00 in the base year (1967 for the United States and 1972 for Japan). Note that these price and implicit quantity indexes satisfy Fisher's weak factor reversal test (Ralph C. Allen and W. Erwin Diewert, 1981).

We evaluate data consistency with various non-parametric hypotheses over the 1958–77 period for the U.S. and 1965–78 period for Japan. Given that these data exhibit zero profit, all of the analyses presented below, except for the separability tests via (5), assume constant return to scale (CRTS) as a maintained hypothesis (see Norsworthy and Malmquist). In the cases where a non-parametric hypothesis is found to be inconsistent with the data, the tests are recomputed after

TABLE 2—RESULTS OF NON-PARAMETRIC TESTS FOR THE WEAK SEPARABILITY OF INPUTS VIA EQUATION (5): U.S. (1958–77) AND JAPANESE (1965–78) MANUFACTURING DATA

Specification/Test	U.S.	Japan
Capital and Labor: $f(K,L)$	Accept	Accept
Materials and Energy: $f(M,E)$	Accept	Accept
Capital, Materials, and Energy: $f(K,M,E)$	Accept	Accept
Capital, Labor, Materials, and Energy: $f(K,L,M,E)$ (Technical Separability or Extended Hicks Neutrality)	Accept	Accept

Source: Computations by the authors.
Note: Accept indicates the existence of a feasible solution to the dual LP system; hence, acceptance of the non-parametric hypothesis test.

including the goodness-of-fit measures as discussed above.

Table 2 summarizes the results of the various inputs weak separability tests. First, we non-parametrically evaluate the weak separability of capital and labor (K,L), materials and energy (M,E), and capital, materials, and energy (K,M,E) via equation (5). We then evaluate the hypothesis of generalized Hicks neutrality as the weak separability of all inputs (K,L,M,E) from output and the technology indexes (A) via equation (5). All input aggregates and the extended Hicks neutrality hypotheses are found to be data consistent over the periods analyzed for both the United States and Japanese manufacturing data. It is insightful to contrast these results with parametric results of Norsworthy and Malmquist obtained with the same data in the context of translog production and unit cost functions. Our non-parametric results indicate the existence of a production function with extended Hicks neutral technical change that is consistent with the United States as well as Japanese data. In contrast, Norsworthy and Malmquist reject Hicks neutral technical change in both their estimated translog production and cost function for the Japanese data.[6] These differences in results likely reflect the Diamond et al. identification issues.

Norsworthy and Malmquist similarly test for the weak separability of capital and labor using a translog production function as well as a unit cost function under alternative maintained hypotheses.[7] They reject this weak separability for both U.S. and Japanese production functions under all maintained hypotheses. They also obtain similar results based on a cost function analysis of Japanese manufacturing (see Norsworthy and Malmquist, p. 953).

For U.S. as well as Japanese manufacturing, our non-parametric results indicate the existence of a production function which rationalizes these data and is weakly separable in (K,L) (as well as (M,E), (K,M,E), and (K,L,M,E)). These results suggest that Norsworthy and Malmquist's rejections of the weak separability of (K,L) using these same data are sensitive to the parametric specification of their model.

Tables 3, 4, and 5 summarize the results of the non-parametric tests for alternative specifications of technical change under the maintained hypotheses of cost minimization and CRTS.[8] As indicated in Table 3, the

[6] For U.S. data, Norsworthy and Malmquist (p. 953) also reject Hicks neutrality using a cost function approach, although they accept the hypothesis of Hicks neutral technical change based on a production function approach.

[7] Their separability test is Michael Denny and Melvyn Fuss's (1978) local separability test (rather than the more restrictive weak global separability of Ernst Berndt and Laurits R. Christensen, 1973).

[8] The goodness-of-fit measures e_t^* are reported in Table 4 only for those hypotheses which were found to be inconsistent with the data (i.e., had unbounded solutions to the dual linear programming formulation of the non-parametric test). In the case where we found data consistency, the associated goodness-of-fit measures are simply $e_t^* = 1$ for all years.

TABLE 3—RESULTS OF NON-PARAMETRIC TESTS FOR ALTERNATIVE SPECIFICATIONS OF TECHNICAL CHANGE UNDER COST MINIMIZATION AND CONSTANT RETURN TO SCALE (CRTS): U.S. (1958–77) AND JAPANESE (1965–78) MANUFACTURING DATA

Specification/Test	U.S.	Japan
Strong Axiom of Cost Min with CRTS: (No Technical Change)	Reject	Reject
CRTS Cost Min: Output Translating (Hicks Neutrality)	Reject	Reject
CRTS Cost Min: Input Translating	Accept	Accept
CRTS Cost Min: Output & Input Translating	Accept	Accept
CRTS Cost Min: Output & Input Translating & Additivity	Reject	Reject

Source: Computations by the authors.
Note: Accept indicates the existence of a feasible solution to the dual LP system; hence, acceptance of the non-parametric hypothesis test. Reject indicates an unbounded solution to the dual LP problem, hence non-parametric rejection of the hypothesis.

TABLE 4—NON-PARAMETRIC GOODNESS-OF-FIT MEASURES FOR ALTERNATIVE TECHNICAL CHANGE SPECIFICATIONS UNDER CRTS COST MINIMIZATION AND/OR ADDITIVE AUGMENTATION: NORSWORTHY AND MALMQUIST U.S. AND JAPANESE MANUFACTURING DATA

	U.S.			Japan		
Year	Strong Axiom of Cost Min	Output Augments	Additivity Input and Output Augments	Strong Axiom of Cost Min	Output Augments	Additivity Input and Output Augments
1958	0.6490	1.0000	0.9931	–	–	–
1959	0.7324	1.0000	0.9976	–	–	–
1960	0.7399	1.0000	0.9978	–	–	–
1961	0.7375	1.0000	0.9960	–	–	–
1962	0.7996	1.0000	0.9980	–	–	–
1963	0.8539	1.0000	0.9998	–	–	–
1964	0.8971	1.0000	1.0000	–	–	–
1965	0.9268	1.0000	0.9997	0.5490	1.0000	0.9593
1966	0.9385	0.9981	0.9983	0.6561	1.0000	0.9833
1967	0.9203	0.9981	0.9989	0.8129	0.9999	1.0000
1968	0.9389	1.0000	0.9991	0.9233	1.0000	1.0003
1969	0.9365	1.0000	0.9984	0.9794	1.0000	0.9994
1970	0.9086	1.0000	0.9990	0.9867	1.0000	0.9984
1971	0.9335	1.0000	0.9994	0.9606	1.0000	0.9995
1972	0.9729	0.9998	0.9964	0.9880	1.0000	0.9993
1973	0.9860	0.9965	0.9928	0.9036	1.0000	0.9916
1974	0.9378	1.0000	0.9974	0.8501	1.0000	0.9939
1975	0.9380	1.0000	1.0000	0.9482	1.0000	1.0007
1976	0.9822	1.0000	1.0000	1.0000	1.0000	0.9985
1977	1.0000	1.0000	1.0000	0.9628	1.0000	0.9979
1978	–	–	–	0.9752	1.0000	0.9898

Source: Computations by the authors using SAS's LP procedure.
Note: These goodness-of-fit measures are reported only for those hypotheses which were found to be inconsistent with these data (i.e., had unbounded solutions to the dual LP formulation of the associated non-parametric inequalities).

strong axiom of cost minimization without technical change via (7) is rejected for both U.S. and Japanese manufacturing. The associated goodness-of-fit measures from Table 4 (Strong Axiom of Cost Min.) indicate that there is rather strong evidence against the null hypothesis of no technical change. These results suggest that technical change has significantly affected U.S. and Japanese manufacturing production technology over the periods analyzed.

Under the maintained hypotheses of CRTS cost minimization, the nature of the technical change is further analyzed in the context of the translating hypotheses discussed in Section II. Under the Hicks neutral output translating hypothesis (additive output augmentation, equation (9a)), consistency is also rejected for both U.S. and Japanese manufacturing for the periods analyzed.[9] The associated goodness-of-fit measures from Table 4, however, suggest that this nonparametric rejection is not strong as only 4 years of the United States (1966, 1967, 1972, 1973) and one year of the Japanese (1967) data required adjustment to achieve data consistency. Given that the magnitude of the goodness-of-fit measures are all very close to 1 for the U.S. as well as Japanese data (see Table 4), the evidence against the null hypothesis appears to be quite weak. For example, measurement errors in the data used for the analysis may be sufficient to explain these goodness-of-fit results. Thus, our results are interpreted here as additional evidence supporting the hypothesis of Hicks neutral technical change.

The input translating hypothesis (additive input augmentations, equation (9b)), is found to be consistent with the data for both countries. This suggests evidence that biased technical change could also rationalize these data. Given this result, it is therefore not surprising that non-parametric evaluation of joint input and output translating hypotheses via (9c) is consistent with these data for both countries (see Table 3).

As discussed in Section III, given an appropriate choice of b in (12),[10] the linear programming results associated with the joint input and output translating hypotheses yield measures of the minimum output augmentations (A_{0t}) in equation (9c). These measures are summarized in Table 5.[11] On average, the output augmentations increase faster for Japanese than U.S. manufacturing. This is consistent with the productivity growth reported by Norsworthy and Malmquist. The A_{0t} for the United States reveal an upward trend from 1958–73 with a major downward year-to-year change in 1970 and several smaller downward year-to-year changes (i.e., 1961, 1963, and 1967). The Japanese A_{0t} reveal an upward trend from 1965–1972 with a downward year to year change in 1971. Both the U.S. and Japanese A_{0t} reveal a major decrease in output augmentation around the 1973 oil price shock. The Japanese A_{0t} show a major decrease starting in 1973 and 1974. In contrast, the U.S. A_{0t} show a major decrease for one year in 1973. These results may reflect differences in the degree of oil import dependence between the United States and Japan where heavier import dependence is reflected by larger and longer adjustments in output augmentation. Also note that the Japanese A_{0t} reflect another year-to-year decrease in 1977 which corresponds to the second oil price shock of the 1970's. The relative magnitudes and duration of the year-to-year decrease, however, is much smaller than in 1973.

[9] Under output scaling (where $Y = y/A_0$) and constant return to scale hypotheses, the corresponding non-parametric test was also conducted. Data consistency for both U.S. and Japanese manufacturing was also rejected under these hypotheses.

[10] The results presented below correspond to $M = 1000$. The findings were found to be insensitive to choosing a larger value for M.

[11] Note that our formulation allows for "regressive" technical change in the sense that $A_{0t} \leq A_{0t-1}$ is permitted (see Table 5). The period following the oil price shocks of the 1970s demonstrates this possibility quite well. Similar patterns of technical change for U.S. and Japanese manufacturing 1965–74 are reported by Dale W. Jorgenson and Mieko Nishimizu (1978). This suggests that large changes in relative prices may adversely affect technological change reflecting the adjustments associated with induced shifts to alternative production processes better adapted to the new economic context.

TABLE 5—ESTIMATED MINIMUM RATES OF OUTPUT AUGMENTATION FROM ADDITIVE OUTPUT AND INPUT AUGMENTATION SPECIFICATION OF TECHNICAL CHANGE UNDER COST MINIMIZATION AND CRTS: NORSWORTHY AND MALMQUIST, U.S. AND JAPANESE MANUFACTURING DATA

	U.S.		Japan	
Year	Output Augments (A_{0t})	Chg in Output Augments	Output Augments (A_{0t})	Chg in Output Augments
1958	0.000	–		
1959	0.042	0.042		
1960	0.047	0.005		
1961	0.039	−0.008		
1962	0.071	0.032		
1963	0.066	−0.005		
1964	0.105	0.040		
1965	0.176	0.071	0.000	–
1966	0.191	0.015	0.020	0.020
1967	0.179	−0.013	0.063	0.043
1968	0.199	0.020	0.110	0.046
1969	0.198	−0.001	0.147	0.037
1970	0.120	−0.078	0.157	0.010
1971	0.198	0.078	0.133	−0.024
1972	0.244	0.045	0.171	0.038
1973	0.262	0.019	0.060	−0.111
1974	0.210	−0.053	0.053	−0.007
1975	0.218	0.008	0.145	0.092
1976	0.251	0.033	0.184	0.039
1977	0.263	0.012	0.152	−0.033
1978	–	–	0.162	0.010

Source: Computations by the authors.
Note: These output augmentation values are consistent with equations (9c) under the input and output translating hypothesis.

Note that the U.S. A_{0t} indicate a strong rebound in output augmentation in 1976, that is, after the 1974–75 U.S. recession. Not until 1977, however, did output augmentation surpass the 1973 (prior highest) levels. In contrast, the Japanese A_{0t} indicate a strong rebound in output augmentation in 1975. By 1976, output augmentation surpassed the 1972 (prior highest) levels. These results appear reasonable and illustrate the usefulness of the non-parametric approach to the measurement of productivity change.

Last, we non-parametrically evaluate the hypothesis of strong (additive) separability of capital and labor in the context of translating. More specifically, we test for the existence of a production function $y = g(x) + A_0 + f(Z)$ via equations (10), where x denotes material and energy while Z denotes (translated) capital and labor. Note that this specification allows additive output (or real value added) augmentation (via A_0) and biased technical change in capital and labor where $Z_i = z_i + A_i$, A_i being the corresponding additive input augmentation. This specification would have to hold for the widespread double-deflation method of calculating (value-added) GNP to be valid (Denny and May).

Table 3 indicates that these data are not consistent with this hypothesis over the time periods analyzed for both countries. This is consistent with the Norsworthy and Malmquist translog findings that a value added aggregate requiring additive separability of (K, L) is inappropriate for either the United States or Japan over the periods analyzed. The associated goodness-of-fit measures in Table 4, however, suggest that the evidence against data consistency is rather weak. That is, despite the numerous years in both series that have goodness-of-fit measures lower than one, the lowest measure is found to be 0.9931 for the United States (1958) and 0.9593 for Japan (1965). These small discrepancies could possibly be explained by data measurement errors alone. If so, Norsworthy and Malmquist's evidence against the value added measure of productivity would be sensitive to parametric specification.

V. Concluding Remarks

This paper presents a non-parametric analysis of technology and technical change in the context of cost minimizing behavior. The non-parametric results of Hanoch and Rothschild, and Varian (1984) are extended to incorporate output augmenting (Hicks neutral) and input augmenting (biased) technical change. In addition, goodness-of-fit measures following Afriat (1967) and Varian (1988) are incorporated to allow further evaluation of the evidence against null hypotheses in non-parametric tests.

The non-parametric approach allows the testing of several hypotheses concerning the existence and nature of technical change which are *less dependent* on parametric specification of production relationships. For example, while Norsworthy and Malmquist found evidence against the hypothesis of Hicks neutral technical change, our non-parametric results indicate the existence of a production function exhibiting Hicks neutral technical change that is consistent with both the U.S. and Japanese data. While we assume cost minimization (i.e., the separability tests of Table 2) as well as constant returns to scale and additive output augmentation (i.e., the output and input translating results of Tables 3 and 4) as maintained hypotheses, our results do not require further parametric specification of the cost or production function. This suggests that Norsworthy and Malmquist's analysis of Hicks neutral technical change is sensitive to their parametric specification.

Alternatively, under CRTS cost minimization and additive augmentation (translating) hypotheses, our non-parametric results found only weak evidence against the additive separability of capital and labor from energy and material in U.S. and Japanese data. This additive separability is required to justify a value-added measurement of productivity (Denny and May). Norsworthy and Malmquist found parametric evidence against this additive separability and argued that productivity measurements should be made based on gross output measures (rather than on value added). Again, our results suggest that Norsworthy and Malmquist's results may be sensitive to their parametric specification.

Given the identification and measurement difficulties confronting parametric approaches to technical change analysis (Diamond et al.), the proposed extensions of the non-parametric approach to include technical change hypotheses provide a powerful heuristic to complement the more traditional parametric approaches. Empirical implementation of these extended non-parametric results can be conveniently performed using standard linear programming algorithms when the resulting inequalities are linear. By illustrating the usefulness of the approach, it is hoped that this paper will stimulate additional research on the non-parametric analysis of economic decisions.

APPENDIX

Proof of Proposition 1:

Consider the saddle-point problem $x^* \geq 0$, $\lambda^* \geq 0$ such that

$$\text{(A1)}\quad L(x,\lambda^*,p,A,y) \geq L(x^*,\lambda^*,p,A,y) \geq L(x^*,\lambda,p,A,y),\ \forall x \geq 0,\ \lambda \geq 0,$$

where $L(x,\lambda,p,A,y) = p'x + \lambda[f(y,A) - g(x,A)]$.

If $g(x,A)$ is strictly increasing and concave in x, then the saddle-point criterion (A1) is a necessary and sufficient condition for x^* to be a global solution to the optimization problem $\text{Min}\{p'x: f(y,A) \leq g(x,A),\ x \geq 0\}$ (see V. A. Sposito, 1975; Samuel Karlin, 1957).

Denote the saddle-point of $L(x,\lambda,p_t,A_t,y_t)$ in (A1) by $\lambda_t = \lambda^*(p_t,A_t,y_t)$ and $x_t = x^*(p_t,A_t,y_t)$. The saddle-point characterization (A1) implies that $\lambda_t \cdot [f(y_t,A_t) - g(x_t,A_t)] = 0$, and

$$L(x_s,\lambda_t,p_t,A_t,y_t) \geq L(x_t,\lambda_t,p_t,A_t,y_t) = p_t'x_t.$$

This yields

$$p_t'x_s + \lambda_t \cdot [f(y_t, A_t) - g(x_s, A_t)] \geq p_t'x_t, \quad \lambda_t \geq 0,$$

Given that $g(x, A)$ is strictly increasing in x, it follows that $\lambda_t > 0$, $f(y_t, A_t) = g(x_t, A_t)$ and

$$\text{(A2)} \quad f(y_t, A_t) - g(x_s, A_t) + \frac{p_t'}{\lambda_t}(x_s - x_t) \geq 0, \quad \lambda_t > 0.$$

This proves (a).

To prove (b), define the function

$$\text{(A3)} \quad G(x_s, A_s) = g(x_s, A_s) + \underset{t}{\text{Min}}\left\{ f(y_t, A_t) - g(x_s, A_t) + \frac{p_t'}{\lambda_t}(x_s - x_t) \right\}.$$

For the case where $s = t$ note that (A2) implies that $\lambda_s \cdot [f(y_s, A_s) - g(x_s, A_s)] \geq 0$. If x_s is feasible, this yields $\lambda_s \cdot [f(y_s, A_s) - g(x_s, A_s)] = 0$, or $f(y_s, A_s) = g(x_s, A_s)$ given $\lambda_s > 0$. Also, in this context, we have $G(x_s, A_s) \leq f(y_s, A_s)$ from (A3) (by definition) while $G(x_s, A_s) \geq g(x_s, A_s)$ from (A2). Given $f(y_s, A_s) = g(x_s, A_s)$, it follows that $G(x_s, A_x) = g(x_s, A_s)$, that is, $G(x_s, A_s)$ gives a representation of $g(x_s, A_s)$ for all data points $s = 1, \dots, T$.

From (A2) and (A3) and given $\Omega = \{x_1, x_2, \dots, x_T\}$, we have:

$$\underset{x}{\text{Min}}\{ p_s'x : f(y_s, A_s) \leq G(x, A_s), x \in \Omega \}$$

$$\geq \min_x \left\{ p_s'x : f(y_s, A_s) \leq g(x, A_s) + f(y_s, A_s) - g(x, A_s) + \frac{p_s'}{\lambda_s}(x - x_s), x \in \Omega \right\}$$

$$= \min_x \{ p_s'x : p_s'x_s \leq p_s'x, x \in \Omega \} = p_s'x_s.$$

It follows that x_s solves $\text{Min}\{ p_s'x : f(y_s, A_s) \leq G(x, A_s), x \in \Omega \}$.

REFERENCES

Afriat, Sydney N., "The Construction of a Utility Function from Expenditure Data," *International Economic Review*, February 1967, *8*, 67–77.

______, "Efficiency Estimates of Production Functions," *International Economic Review*, October 1972, *13*, 568–98.

Allen, Ralph and Diewert, W. Erwin, "Direct Versus Indirect Implicit Superlative Index Number Formulas," *Review of Economics and Statistics*, August 1981, *63*, 430–35.

Berndt, Ernest and Christensen, Laurits, R., "The Specification of Technology in U.S. Manufacturing," Working Paper No. 18, Bureau of Labor Statistics, Washington, 1973.

Binswanger, Hans Peter, "The Measurement of Technical Change Bias with Many Factors of Production," *American Economic Review*, December 1974, *64*, 964–76.

Blackorby, Charles, Lovell, C. A. Knox and Thursby, Marie C., "Extended Hicks Neutral Technical Change," *Economic Journal*, December 1976, *86*, 845–52.

Denny, Michael and May, J. Douglas, "Homotheticity and Real Value Added in Canadian Manufacturing," in Melvyn Fuss and Daniel McFadden, eds., *Production Economics: A Dual Approach to Theory and Applications*, Vol. 2, Amsterdam: North-Holland, 1978, ch. III.3.

______ **and Fuss, Melvyn,** "The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates," *American Economic Review*, June 1977, *67*, 404–18.

Diamond, Peter, McFadden, Daniel and Rodriguez, Miguel, "Measurement of the Elasticity of Factor Substitution and Bias of Technical Change," in Melvyn Fuss and Daniel McFadden, eds., *Production Economics: A Dual Approach to Theory and Application*, Vol. 2, Amsterdam: North-Holland, 1978, ch. IV.2.

Griliches, Zvi, "The Demand for Fertilizer: An Economic Interpretation of a Technical Change," *Journal of Farm Economics*, August 1958, *40*, 591–606.

Hanoch, Giora and Rothschild, Michael, "Testing the Assumptions of Production Theory: A Nonparametric Approach," *Journal of Political Economy*, March-April 1972, *80*, 256–75.

Jorgenson, Dale W. and Griliches, Zvi, "The Explanation of Productivity Change," *Review of Economic Studies*, July 1967, *34*, 249–83.

______ **and Nishimizu, Mieko,** "U.S. and

Japanese Economic Growth, 1952–1974: An International Comparison," *Economic Journal*, December 1978, *88*, 707–26.

Karlin, Samuel, *Mathematical Methods and Theory in Games, Programming and Economics*, Vol. I, Palo Alto, CA: Addison Wesley, 1959.

Luenberger, David G., *Linear and Nonlinear Programming*, 2nd ed., Reading, MA: Addison Wesley, 1984.

Norsworthy, John R. and Malmquist, David H., "Input Measurement and Productivity Growth in Japanese and U.S. Manufacturing," *American Economic Review*, December 1983, *73*, 947–67.

Sato, Ryuzo, "The Estimation of Biased Technical Progress and the Production Function," *International Economic Review*, June 1970, *11*, 179–208.

Slater, M., "Lagrange Multipliers Revisited: A Contribution to Nonlinear Programming," Rand Corporation Report RM-676, Santa Monica, CA, 1951.

Sposito, V. A., *Linear and Nonlinear Programming*, Ames, IA: Iowa State University Press, 1975.

Stevenson, Rodney E., "Measuring Technological Bias," *American Economic Review*, March 1980, *70*, 162–73.

Stigler, George J., "Economic Problems in Measuring Changes in Productivity," *Output, Input and Productivity Measurement: Studies in Income and Wealth*. NBER, Princeton: Princeton University Press, 1961, 47–63.

Varian, Hal R., "The Nonparametric Approach to Production Analysis," *Econometrica*, May 1984, *52*, 579–97.

______, "Goodness-of-Fit in Demand Analysis," CREST Working Paper, Department of Economics, University of Michigan, Ann Arbor, September 1988.

Herd Behavior and Investment

By DAVID S. SCHARFSTEIN AND JEREMY C. STEIN*

This paper examines some of the forces that can lead to herd behavior in investment. Under certain circumstances, managers simply mimic the investment decisions of other managers, ignoring substantive private information. Although this behavior is inefficient from a social standpoint, it can be rational from the perspective of managers who are concerned about their reputations in the labor market. We discuss applications of the model to corporate investment, the stock market, and decision making within firms. (JEL 026, 522)

A basic tenet of classical economic theory is that investment decisions reflect agents' rationally formed expectations; decisions are made using all available information in an efficient manner. A contrasting view is that investment is also driven by group psychology, which weakens the link between information and market outcomes. In *The General Theory*, John Maynard Keynes (1936, pp. 157–58) expresses skepticism about the ability and inclination of "long-term investors" to buck market trends and ensure efficient investment. In his view, investors may be reluctant to act according to their own information and beliefs, fearing that their contrarian behavior will damage their reputations as sensible decision makers:

> ...it is the long-term investor, he who most promotes the public interest, who will in practice come in for most criticism, wherever investment funds are managed by committees or boards or banks. For it is in the essence of his behavior that he should be eccentric, unconventional, and rash in the eyes of average opinion. If he is successful, that will only confirm the general belief in his rashness; and if in the short-run he is unsuccessful, which is very likely, he will not receive much mercy. Worldly wisdom teaches that it is better for reputation to fail conventionally than to succeed unconventionally.

Thus Keynes suggests that professional managers will "follow the herd" if they are concerned about how others will assess their ability to make sound judgments. There are a number of settings in which this kind of herd behavior might have important implications. One example is the stock market, for which the following explanation of the pre-October 1987 bull market is often repeated: The consensus among professional money managers was that price levels were too high —the market was, in their opinion, more likely to go down rather than up. However, few money managers were eager to sell their equity holdings. If the market did continue to go up, they were afraid of being perceived as lone fools for missing out on the ride. On the other hand, in the more likely event of a market decline, there would be comfort in numbers—how bad could they look if everybody else had suffered the same fate?

The same principle can apply to corporate investment, when a number of companies are investing in similar assets. In *Selling Money*, Samuel Gwynne (1986, p. 58) documents problems of herd behavior in banks' lending policies toward LDCs. Discussing the incentives facing a credit analyst, he

*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139 and Harvard Business School, Boston, MA 02163. We are grateful for helpful comments from Robert Gertner, Bob Gibbons, Bengt Holmstrom, Chi-Fu Huang, Robert Merton, Jim Poterba, Julio Rotemberg, Garth Saloner, Andrei Shleifer, Larry Summers, Robert Vishny, Mark Wolfson, and two anonymous referees, and participants in seminars at the NBER Summer Institute, Harvard, Rochester, Columbia, and Northwestern.

writes:

> Part of Herrick's job—an extremely important part as far as the bank was concerned—was to retrieve information about the countries in which the bank did business. But this function collided head on with what Herrick was actually doing out there...His job would never be measured by how correct his country risk analysis was. At the very least, Herrick was simply doing what hundreds of other larger international banks had already done, and any ultimate blame for poor forecasting would be shared by tens of thousands of bankers around the world; this was one of the curious benefits of following the herd.

The aim of this paper is to develop a clearer understanding of some of the forces that can lead to herd behavior. We find that, under certain circumstances, managers simply mimic the investment decisions of other managers, ignoring substantive private information. Although this behavior is inefficient from a social standpoint, it can be rational from the perspective of managers who are concerned about their reputations in the labor market.

Our model is a "learning" model, similar in spirit to one studied by Bengt Holmstrom (1982a). Like us, he considers a situation in which managers use investment decisions to manipulate the labor market's inferences regarding their ability, where ability represents an aptitude for making decisions. This definition of ability contrasts with one where ability adds to physical productivity, as in Holmstrom and Joan Ricart i Costa (1986) as well as in another part of Holmstrom (1982a). The key difference between our model and these others is that ours is most interesting when there is more than one manager, whereas theirs are single-manager models.[1]

We assume that there are two types of managers: "smart" ones, who receive informative signals about the value of an investment, and "dumb" ones, who receive purely noisy signals. Initially, neither the managers themselves nor the labor market can identify the types. However, after the managers have made an investment decision, the labor market can update its beliefs, based on two pieces of evidence: 1) whether the manager made a profitable investment; 2) whether the manager's behavior was similar to or different from that of other managers.

If there are systematically unpredictable components of investment value, the first piece of evidence will not be used exclusively, since on any given draw, all smart managers could get unlucky and receive misleading signals. Hence the second piece of evidence is important as well. Holding the absolute profitability of the investment choice fixed, managers will be more favorably evaluated if they follow the decisions of others than if they behave in a contrarian fashion. Thus an unprofitable decision is not as bad for reputation when others make the same mistake—they can share the blame if there are systematically unpredictable shocks.

This "sharing-the-blame" effect arises because smart managers tend to receive correlated signals (since they are all observing a piece of the same "truth"), while dumb ones do not (they simply observe uncorrelated noise). Consequently, if one manager mimics the behavior of others, this suggests to the labor market that he has received a signal that is correlated with theirs, and is more likely to be smart. In contrast, a manager who takes a contrarian position is perceived as more likely to be dumb, all else being equal. Thus even if a manager's private information tells him that an investment has a negative expected value, he may pursue it if others before him have. Conversely, he may refuse investments that he perceives as having positive expected value if others before him have also done so.

[1] In Holmstrom's (1982a) example where talent is related to the ability to make good decisions, there are inefficiencies with only one manager. This follows from his assumption that the outcome of a potential investment project is unobservable when the investment is not undertaken. Our model differs in that the state of the world that determines investment profitability is always observable. In the case that we study, there are thus no inefficiencies in a single-manager setting.

These points are further developed in the five sections following this one. In Section I, we present the structure and assumptions of the model. The basic results on the existence of herding equilibria are summarized in three propositions in Section II. Section III looks more closely at some countervailing forces that may offset herding tendencies. In Section IV, we elaborate on the implications of the model for corporate investment, the stock market, and decision making within firms. Finally, Section V contains concluding remarks.

I. The Model

The model that is developed in this section applies more literally to the example of corporate investment discussed above than it does to the stock market. We assume that the investments under consideration are available in perfectly elastic supply at a given price. This allows us to avoid explicitly considering the feedback from investment demand to prices, thereby simplifying the analysis considerably. In Section IV, we will discuss at greater length how we think our results carry over to the stock market, where this assumption is clearly not appropriate. For the moment, however, it may help for the reader interested in concreteness to bear in mind the following story: In our model, managers are in charge of capital investment at industrial firms, and they each are considering investing in a cost-saving technology. The value of the technology will be realized in the future, and each manager receives a signal that gives him some information about the future state. The question we address is the following: How well does the aggregate level of investment reflect all of the available information?

A. *Timing and Information Structure*

The economy consists of two firms, firm A and firm B, run by managers we call A and B, respectively. These managers invest sequentially, with A moving first. At date 1, A decides whether or not to make the investment. There are two possible outcomes at date 3: either the "high" state, in which case the investment yields a profit (net of investment expense and discounting) of $x_H > 0$; or the "low" state, in which case the net profit is $x_L < 0$. The prior probabilities of these two states are α and $(1-\alpha)$, respectively. The outcome is publicly observable, even if neither manager decides to invest.

In making his decision, A has access to a signal, which can take on one of two values: s_G (a "good" signal); or s_B (a "bad" signal). Interpreting this signal is a bit complicated, because the manager does not know if he is "smart" or "dumb." If he is smart, which occurs with prior probability θ, the signal is informative—that is, a good signal is more likely to occur prior to the high state than the low state. Formally, we have:

$$\text{(1)} \qquad \text{Prob}(s_G|x_H, \text{smart}) \equiv p;$$

$$\text{(2)} \qquad \text{Prob}(s_G|x_L, \text{smart}) \equiv q < p.$$

If the manager is dumb, however, which occurs with probability $(1-\theta)$, he receives completely uninformative signals—he is as likely to receive s_G prior to the high state as prior to the low state:

$$\text{(3)} \quad \text{Prob}(s_G|x_H, \text{dumb}) = \text{Prob}(s_G|x_L, \text{dumb}) \equiv z.$$

We make the assumption that the *ex ante* distribution of signals is the same for both smart and dumb managers—both are equally likely to receive s_G, so that the actual signal received does not communicate any information about the manager's type. This amounts to assuming that: $\text{Prob}(s_G|\text{smart}) = \text{Prob}(s_G|\text{dumb})$, or:

$$\text{(4)} \qquad z = \alpha p + (1-\alpha) q.$$

Given that the manager does not know if he is smart or dumb, a straightforward application of Bayes' law allows us to calculate the probabilities he attaches to the high state after receiving the good and bad signals:

$$(5)\quad \mathrm{Prob}(x_H|s_G) \equiv \mu_G = \frac{[\theta p + (1-\theta)z]}{z}\alpha;$$

$$(6)\quad \mathrm{Prob}(x_H|s_B) \equiv \mu_B = \frac{[\theta(1-p)+(1-\theta)(1-z)]}{(1-z)}\alpha.$$

In order to make the investment problem interesting, we assume that the investment is attractive if a good signal has been received, but not if a bad signal has been received:

$$(7)\quad \mu_G x_H + (1-\mu_G)x_L > 0 > \mu_B x_H + (1-\mu_B)x_L.$$

After A has made his investment decision at date 1, B makes his decision at date 2. Manager B also has access to a private signal. In addition, he can observe whether or not firm A has decided to invest. This is valuable information; even in a world without reputational concerns, firm B's investment decision should be partially influenced by what firm A does. Our main point is that with reputational concerns, firm B's manager pays *too much* attention to what firm A has done, and too little to his private signal.[2]

Like manager A, firm B's manager can be either smart or dumb. If one manager is smart and the other is dumb, their signals are drawn independently from the binomial distributions given in equations (1)–(3). Similarly, if both are dumb, their signals are drawn independently—so that, for example, the probability is z^2 that two dumb managers both observe the good signal.[3]

However, if both managers are smart, they are assumed to observe *exactly the same signal.* Thus the probability that two smart managers both observe s_G when the true state is x_H is p (as opposed to p^2 if smart managers received independent draws from the distributions described in (1) and (2)). This feature is crucial to our analysis. It can be generalized somewhat to allow the draws to be imperfectly correlated. However, if the signals of smart managers are drawn independently from the distributions, our results concerning herd behavior fail to go through.

Heuristically, herd behavior requires smart managers' prediction errors to be at least partially correlated with each other. Although this feature may seem slightly unnatural given the current setup and notation, it amounts to nothing more than saying that there are *systematically unpredictable factors affecting the future state that nobody can know anything about.*[4] For example, we might model the outcome of the state draw as being driven by the sum of two random variables, u and v. If $u+v>0$, then x_H obtains. If $u+v<0$, then x_L obtains. Our assumption is equivalent to allowing smart

[2] An analogy to noisy rational expectations models of the stock market (such as Martin Hellwig, 1980), where all the players move simultaneously, may be helpful. In these models, traders also put some weight on information drawn from the investment decisions of others (which are reflected in the stock price) and some weight on their private information. The question we address here is whether too little weight is put on private information in making decisions. Note, however, that in a noiseless stock market, the price reveals all information, so that ignoring private signals is actually optimal. See, for example, Paul Milgrom and Nancy Stokey (1982).

[3] Our assumption that all dumb managers receive independent signals may seem extreme. Many investors who respond to the *same* stimuli—for example, the predictions of technical stock market forecasters—are often thought to be "dumb" for doing so. However, care must be taken in interpreting this fact with respect to our model. We only assume that dumb agents' *private* signals are independent—that is, these agents make independent errors when trying to interpret data *on their own*. The fact that people rely on common *outside* sources for forecasts in a suboptimal fashion is not at all inconsistent with our approach. Indeed, it is essentially our main prediction: privately formed opinions will be disregarded in favor of publicly available forecasts.

[4] Empirical evidence suggest that the assumption of systematically unpredictable components is a reasonable one. For example, Patricia O'Brien (1988) finds that individual security analysts' prediction errors contain common components.

managers to observe u but not v. (As suggested above, we could generalize to allow each manager to observe u perturbed by independent noise, so long as we retained the unobservability of the common component v.)

The importance of a common component to the prediction error for smart managers follows logic similar to that seen in effort-based models of relative performance such as in Holmstrom (1982b) and Barry Nalebuff and Joseph Stiglitz (1983). If prediction errors are independent, the labor market can, in our model, efficiently update on ability using only individual performance—that is, whether the manager picked a successful investment. Analogously, in models of relative performance, if the agents face independent shocks to output, optimal contracts evaluate the agent based only on individual absolute performance. However, in these models, if errors are correlated, there is an informational gain from comparing agents. In our model, one will not wish to evaluate too harshly a manager who picks a bad investment, if his colleague's similar choice suggests that they were both victims of a completely unpredictable factor. This is the "sharing the blame" effect touched on earlier.

Thus it is the common component to prediction errors that gives this model its bite, by causing some inferential weight to be placed on the similarity of managers' decisions. Perversely, the existence of this extra channel of inference actually leads to *ex ante* reductions in efficiency—just the opposite of the result seen in the literature on tournaments. This is because here, managers attempt to actively manipulate their investment decisions in such a way as to bias the inference process in their favor. Even if the market recognizes that they will be engaging in this manipulation, it will continue to exist in equilibrium.[5]

B. *Managerial Objectives*

Our next step is to specify the managers' objective functions. In a first-best world, the managers would seek only to maximize the expected returns on investment, and would invest anytime their information (either from their private signal or from observing the other managers) indicated that investing had positive expected value.

In our model, managers' investment decisions enable the labor market to update its beliefs about their ability. We denote by $\hat{\theta}$ the market's revised assessment of the probability that a manager is smart.

In order to establish a simple relationship between managers' objectives and $\hat{\theta}$, we make several simplifying assumptions. Following Holmstrom and Ricart i Costa (1986), we assume that: 1) the investment game is replayed once more after date 3; 2) at this point there is no further reason to build a reputation, so managers invest efficiently; 3) competition leads managers' spot market wages to be set to the economic value of their ability.

It is straightforward to demonstrate that, for a wide range of parameter values in our model, the expected return on the investment opportunity is linear in the manager's ability (as measured by $\hat{\theta}$) if the manager invests efficiently. Hence the spot market wages referred to above will be proportional to $\hat{\theta}$.[6]

We do not explicitly analyze contracting behavior in what follows. Rather, we assume (as do Holmstrom, 1982a, and others) that managers cannot be bound to their firms against their will *ex post*. This means that any long-term contract that would pay some types less than spot market wages in the second round of the investment game is infeasible.

Since their future wages are linear in $\hat{\theta}$, managers have some incentive to generate

[5] The basic idea about manipulation of the learning process was first developed by Holmstrom (1982a). Drew Fudenberg and Jean Tirole (1986) apply the concept and refer to it as "signal jamming."

[6] This will be the case if the manager's investment decision in the second go-round of the investment does not depend on θ; the manager invests if and only if he observes the good signal. This condition is analogous to that posited for the first go-round in inequality (7).

high values of $\hat{\theta}$, rather than to invest efficiently in the first round. Of course, it is still possible that short-term incentive contracts could serve at least partially to align managerial and firm interests, by specifying a profit-contingent wage in the first round of the investment game. Thus, in principle, it seems reasonable to believe that managers would be induced to act so as to maximize a weighted average of expected profits and their future compensation. However, this more general formulation leads to the same basic conclusions that obtain if managers care *only* about reputation—although naturally, more weight on expected profits will tend to attenuate the inefficiencies. For the sake of starkness and notational simplicity, we leave expected profits out of the managerial objective function. Later, in discussing our results we briefly touch on how they would be altered if managers cared about expected profits.[7]

The last assumption we make is that managers are risk neutral, so that their objective function simplifies to maximizing expected wages. This is equivalent to maximizing the expected value of $\hat{\theta}$.

It should be noted that managers care only about their *absolute* ability assessment —not about whether they are judged to be more or less able than other managers. We touch on the importance of relative ability in Section III.

II. Herding Equilibria

A. *Comparison with Efficient Investment Decisions*

In order to economize on notation, we set $p = 1 - q$. We also set $\alpha = \frac{1}{2}$. Taken together, these simplifications imply that $z = \frac{1}{2}$ from equation (4). For now, we leave the sign of $(x_H + x_L)$ unspecified, so that the investment can have an *ex ante* expected value that is either positive or negative.

As a benchmark, we first derive the optimal decision rules in a first-best world with no reputational concerns. Manager A would invest if and only if he observed s_G—this is a consequence of the assumption in equation (7). Thus manager B can infer manager A's signal from his investment decision.

If manager B observes s_B after firm A has invested, he makes his decision based on the two-signal information set (s_G, s_B). Given our symmetry assumptions, this implies the probability of the high state, $\text{prob}(x_H | s_G, s_B) = \frac{1}{2}$. Thus the investment decision hinges on the sign of $(x_H + x_L)$—if this quantity is positive, manager B will invest, and if not, he will not.

Similarly, if firm A does not invest, and manager B observes s_G, the investment decision turns on the same criterion of whether $(x_H + x_L) > 0$. Clearly, in the first best, the order in which the information arrives is irrelevant to manager B's decision. If one manager observes s_G and the other sees s_B, manager B's decision will be the same regardless of whether the s_G signal was received by him or by manager A.

With reputational considerations, the decision rules are different. When manager A observes s_G and invests, manager B will also invest, regardless of his signal and the sign of $(x_H + x_L)$. Hence if this signal is s_B and $x_H + x_L < 0$, the investment will be inefficient. Conversely, if firm A does not invest, firm B never will either, which is inefficient when manager B observes s_G and $x_H + x_L > 0$. Now the order in which information arrives is important to firm B's decision—the same aggregate information of (s_G, s_B) can lead it to invest or not to invest, depending on whether the signal s_G is received by the first mover firm A or not.

B. *Equilibria with Reputational Concerns*

We now examine the equilibria that exist when managers seek to maximize the expected value of $\hat{\theta}$. For now, we focus on the decision rules of manager B—in all the "continuation" equilibria that we look at the manager A behaves efficiently, by investing

[7] Our simplified formulation would be most literally applicable to situations where the state is publicly observable, but cannot be verified by the courts, so that profit-contingent contracts are not feasible. (For more discussion of this point, see Oliver Hart and Holmstrom, 1987.) This may be a reasonable assumption for certain types of jobs where there are no easily describable performance measures, but it is not valid in other cases, for example, portfolio management.

if and only if he observes s_G. Later, we will establish that this efficient behavior by firm A's manager is part of an equilibrium of the overall game.

We develop our results through a series of propositions.

PROPOSITION 1: *There does not exist any continuation equilibrium in which manager B's investment decision depends on the signal he observes. Thus the only possible equilibria are those where manager B mimics manager A regardless of the signal, or where manager B does the opposite of manager A regardless of the signal.*

The proof will be by contradiction. We start by conjecturing the existence of the "separating" equilibrium described above.[8] We then determine the updating rules the labor market would use to calculate $\hat{\theta}$ in such an equilibrium. Finally, we show that given these updating rules, rational managers will not wish to behave as posited in the equilibrium.

The revised ability assessments will be a function of the labor market's conjectures about the signals observed by the managers as well as the realized state of the world. Of course, only the managers observe their signals, but in the putative separating equilibrium, there is a one-to-one mapping from signals to actions. Thus for example, suppose the separating equilibrium calls for each manager to invest if and only if he observes s_G. Then if manager A does not invest and manager B does, the market believes that manager A observed s_B and manager B observed s_G, and it can do its updating based on these beliefs.

As noted above, the main focus of our analysis is manager B. Continuing with the above example, suppose the high state was realized. How would the market revise its prior about the manager's ability? Let (s_B, s_G, x_H) denote this event, and let $\hat{\theta}(s_B, s_G, x_H)$ be the revised prior. By Bayes' rule, one can show that:

$$(8)\quad \hat{\theta}(s_B, s_G, x_H) = \frac{\frac{1}{2}p\theta(1-\theta)}{\frac{1}{2}p\theta(1-\theta)+\frac{1}{2}(1-p)\theta(1-\theta)+\frac{1}{4}(1-\theta)^2} = 2\theta p/(1+\theta).$$

The explanation for this result is as follows. There are three possible configurations of managerial ability that could give rise to this event: (dumb, smart); (smart, dumb); and (dumb, dumb). Note that (smart, smart) is not possible since if this were the case, both managers would have received the same signal, by virtue of our assumption that smart managers' signals are perfectly correlated.

We wish to know the probability of (dumb, smart) conditional on the event (s_B, s_G, x_H). If the configuration of talent is (dumb, smart), which occurs with *ex ante* probability $\theta(1-\theta)$, the probability of (s_B, s_G) in state x_H is $\frac{1}{2}p$. This explains the numerator in (8). The denominator gives in addition the probabilities of $(s_B, s_G|x_H)$ if the configuration is (smart, dumb) and (dumb, dumb). These are $\frac{1}{2}(1-p)$ and $\frac{1}{4}$, respectively. By symmetry, it is straightforward to show that $\theta(s_G, s_B, x_L) = 2\theta p/(1+\theta)$ also.

The derivation of the other updating rules follow along similar lines. They are listed below:

$$(9)\quad \hat{\theta}(s_B, s_G, x_L) = \hat{\theta}(s_G, s_B, x_H) = 2\theta(1-p)/(1+\theta);$$

$$(10)\quad \hat{\theta}(s_B, s_B, x_H) = \hat{\theta}(s_G, s_G, x_L) = \frac{2\theta(1-p)(1+\theta)}{4\theta(1-p)+(1-\theta)^2};$$

$$(11)\quad \hat{\theta}(s_B, s_B, x_L) = \hat{\theta}(s_G, s_G, x_H) = \frac{2\theta p(1+\theta)}{4\theta p+(1-\theta)^2}.$$

[8] It is important to keep in mind that the only private information of the manager is about the signal he observes, not about his ability. Hence the "separation" is with respect to this signal.

Now, in order for our posited equilibrium to actually hold together, it must be that managers find it in their interest to behave as assumed. For example, suppose that manager A has observed s_B and has not invested. Given the updating rules above, can it ever be rational for manager B to invest upon observing s_G, but not invest upon observing s_B? Or will one type of manager "break" the equilibrium by deviating and attempting to fool the market into thinking that he has received a different signal?

To answer these questions, the following probabilities must be calculated:

$$\text{(12)} \qquad \text{Prob}(x_H|s_B, s_G) = \tfrac{1}{2};$$

$$\text{(13)} \quad \text{Prob}(x_H|s_B, s_B) = \frac{4\theta(1-p)+(1-\theta)^2}{4\theta+2(1-\theta)^2}.$$

We can now check the rationality conditions that must hold for the equilibrium to be viable. One of these is

$$\text{(14)} \quad \hat{\theta}(s_B, s_G, x_H)\text{Prob}(x_H|s_B, s_G) + \hat{\theta}(s_B, s_G, x_L)\text{Prob}(x_L|s_B, s_G) \geq \hat{\theta}(s_B, s_B, x_H)\text{Prob}(x_H|s_B, s_G) + \hat{\theta}(s_B, s_B, x_L)\text{Prob}(x_L|s_B, s_G).$$

Inequality (14) represents the requirement that if manager B receives signal s_G, he prefers to invest (and identify himself as someone who had observed s_G), rather than not invest (and masquerade as someone who had received signal s_B). Direct substitution from equations (8)–(12) establishes that the inequality is violated—if manager B receives s_G, he will wish to deviate by mimicking firm A and not investing. A symmetric argument establishes that if firm A *has* invested, there is also no separating equilibrium. In this case, if manager B observes s_B, he will deviate by mimicking firm A and also investing.

There are also potentially "perverse" separating equilibria, where manager B invests if and only if he observes s_B, rather than s_G. It can be easily demonstrated using the same lines of reasoning that such equilibria are also not viable. This completes the proof of Proposition 1.

It is worth examining the updating rules in equations (8)–(11) to gain some intuition for the forces that break the efficient equilibrium. Two main points emerge from these equations:

First, $\hat{\theta}(s_B, s_G, x_H) > \hat{\theta}(s_B, s_G, x_L)$; and $\hat{\theta}(s_G, s_G, x_H) > \hat{\theta}(s_G, s_G, x_L)$. Holding the investment decision of manager A fixed, manager B is indeed compensated for making "absolutely" good decisions—for investing prior to a realization of x_H, as opposed to investing prior to a realization of x_L.

Second, however, the investment decision of manager A does have an important externality effect. Holding the correctness of the investment decision fixed, there is a higher payoff to manager B for imitating manager A. That is, $\hat{\theta}(s_G, s_G, x_H) > \hat{\theta}(s_B, s_G, x_H)$; and $\hat{\theta}(s_G, s_G, x_L) > \hat{\theta}(s_B, s_G, x_L)$.

Proposition 1 is a direct consequence of this second effect. Because of the payoff to imitation, even if the new information makes it more likely that contradicting manager A is the economically correct decision, manager B prefers to mimic A. As a result, decisions cannot be made contingent on signals, and there cannot be an equilibrium where manager B takes advantage of his private information.

As was emphasized in the previous section, the result depends on our assumption that prediction errors are correlated across smart managers. If the signals of smart managers are independent, Proposition 1 no longer holds. This can be demonstrated by calculating the updating rules that would prevail if signals were independent. Denoting these rules by $\hat{\theta}^i(\ )$, and using the same Bayesian logic as before, we can derive

$$\text{(15)} \quad \hat{\theta}^i(s_B, s_G, x_H) = \hat{\theta}^i(s_G, s_G, x_H) = \hat{\theta}^i(s_B, s_B, x_L) = \hat{\theta}^i(s_G, s_B, x_L) = \frac{2\theta p}{2\theta p+(1-\theta)};$$

$$(16)\quad \hat{\theta}^i(s_B, s_G, x_L) = \hat{\theta}^i(s_G, s_G, x_L) = \hat{\theta}^i(s_B, s_B, x_H) = \hat{\theta}^i(s_G, s_B, x_H) = \frac{2\theta(1-p)}{2\theta(1-p)+(1-\theta)}.$$

According to equations (15) and (16), in the case of independent signals, the labor market's assessment of firm B's manager is unrelated to the investment decision of firm A. All that matters is the profitability of the investment—investing before state x_H leads to a more favorable ability assessment than not investing before state x_H. As a result, the inequality in (14) is satisfied (with equality) and it is possible to sustain the efficient equilibrium in which manager B's investment decisions generally depend on his private signal.

If the independence assumption is relaxed, the updating rules become a function of firm A's investment decision, inequality (14) is violated, and Proposition 1 holds. Thus perfect correlation of smart manager signals is not necessary for our results—all that is needed is *some* correlation of prediction errors.

Having established that continuation equilibria with signal-contingent decisions by manager B do not exist, we now turn our attention to the equilibria that can be supported in our model.

PROPOSITION 2: *There exists a continuation equilibrium in which manager B always mimics manager A, investing if and only if A does. This herding equilibrium is supported by the following "reasonable" out of equilibrium beliefs: i) if manager B deviates by investing when A has not, the labor market believes that he observed s_G; and conversely, ii) if manager B deviates by not investing when A has, the labor market believes that he observed s_B.*

In order to prove Proposition 2, it is necessary to show that manager B will always find it optimal to behave as prescribed, given the beliefs posited. Let us consider only the case where firm A has already not invested; the other case works exactly the same way.

If manager B follows firm A by also not investing, his revised ability assessment is simply equal to θ—there is no revision from the prior because the equilibrium is a "pooling" one, with both s_G and s_B recipients choosing the same action.

In order for manager B who observes s_B not to deviate, it must be the case that:

$$(17)\quad \theta \geq \hat{\theta}(s_B, s_G, x_H)\text{Prob}(x_H|s_B, s_B) + \hat{\theta}(s_B, s_G, x_L)\text{Prob}(x_L|s_B, s_B).$$

Inequality (17) is the requirement that the payoff to a manager who observes s_B, pools, and receives θ, exceeds his payoff from deviating and investing, given that out of equilibrium beliefs are such that he will be viewed as having observed s_G if he deviates. Direct substitution from (8), (9), and (13) verifies that the inequality is satisfied.

In order for manager B observing s_G not to deviate, the following must hold:

$$(18)\quad \theta \geq \hat{\theta}(s_B, s_G, x_H)\text{Prob}(x_H|s_B, s_G) + \hat{\theta}(s_B, s_G, x_L)\text{Prob}(x_L|s_B, s_G).$$

Comparison of (17) and (18) shows that it is relatively more tempting for manager B to deviate by investing after observing s_G, as opposed to s_B. (It is in this sense that the out-of-equilibrium conjecture that a deviator has seen s_G is "reasonable.") Nonetheless, (18) reduces to

$$(18')\quad \theta \geq \theta/(1+\theta),$$

which is strictly satisfied. Thus Proposition 2 is proved, and we have established the existence of a herding continuation equilibrium.

It should be pointed out that there is another, perverse continuation equilibrium in which the decisions of manager B do not depend on his signal. In this equilibrium, manager B always contradicts manager A, investing if and only if A has not. This equilibrium can only be supported by the following "unreasonable" beliefs off the equilibrium path; if manager B deviates by investing when the equilibrium calls for him not to, it is because he has observed s_B; and if he deviates by not investing when the equilibrium requires investment, it is because he has observed s_G.

The multiplicity of equilibria stems from our assumption that investment decisions do not directly affect the manager's utility, but are nothing more than a means of conveying information to the labor market about the signal that the manager has observed. If the labor market (perversely) interprets investment to mean that the manager has observed s_B, then the manager may well invest if he wishes to convince the labor market that he has seen s_B. It follows that in the current formulation of our model, we cannot pin down exactly what actions will be taken. However, we can pin down how much information is revealed in equilibrium: In both equilibria, managers' actions do not depend on their private signals and hence convey no information.[9]

This reasoning suggests that the model can be altered slightly so as to leave the herding equilibrium as the unique outcome. Suppose that managers do not invest directly; instead, they report their signals to the "owners" of the firm, who then make investment decisions to maximize profits. Proposition 1 then can be interpreted as saying that there is no equilibrium where manager B can be relied on to make informative reports. Given that he cannot learn anything from his manager, the owner of firm B will then have to rely on the only available information, the action of firm A. The unique profit-maximizing decision for the owner of firm B is thus always to mimic firm A.

In sum, then, the contradiction equilibrium is probably not a sensible one. If one dismisses it, the herding equilibrium is left as the unique continuation equilibrium of the game. It remains only to establish that the efficient behavior on the part of manager A that has been assumed to this point is part of an overall equilibrium.

PROPOSITION 3: *There exists an equilibrium for the overall game where manager A invests if and only if he receives s_G, and where manager B always mimics manager A regardless of B's signal.*

In the proposed equilibrium, there is no information inherent in manager B's actions, since he always does the same thing, regardless of his signal. Thus manager A can only be evaluated absolutely—his revised ability $\hat{\theta}^A$ is a function of only his action and the realized state. The updating rules are therefore identical to those given for the two-manager, independent signal case in equations (15) and (16). That is

$$(19)\quad \hat{\theta}^A(s_G, x_H) = \hat{\theta}^A(s_B, x_L) = 2\theta p/(2\theta p + (1-\theta)); \text{ and}$$

$$(20)\quad \hat{\theta}^A(s_G, x_L) = \hat{\theta}^A(s_B, x_H) = 2\theta(1-p)/(2\theta(1-p) + (1-\theta)).$$

In order for manager A to be willing to invest after observing s_G and not invest after observing s_B, the following two conditions

[9] The structure of our model is similar in some respects to Vincent Crawford and Joel Sobel's (1982) "cheap talk" model of strategic information transmission. In their framework, an informed party (the "sender") transmits a message to another party (the "receiver"), who then takes an action that affects the utility of both parties. Similarly, in our model, one can view the investment decision as a message sent by the informed manager and the wage paid by the outside labor market as the response of the receiver. (However, our model differs from theirs in that the message itself —the investment decision—has real economic consequences to a third party, namely, the manager's firm.) What matters in these models is not the actual wording of the messages, but rather how they are interpreted by the receiver. Hence, it is impossible to determine exactly what words will be sent, though it is possible to determine how much information is contained in these words.

must hold:

$$(21)\quad \hat{\theta}^A(s_G, x_H)\mu_G + \hat{\theta}^A(s_G, x_L)(1-\mu_G) \geq \hat{\theta}^A(s_B, x_H)\mu_G + \hat{\theta}^A(s_B, x_L)(1-\mu_G);$$

$$(22)\quad \hat{\theta}^A(s_B, x_H)\mu_B + \hat{\theta}^A(s_B, x_L)(1-\mu_B) \geq \hat{\theta}^A(s_G, x_H)\mu_B + \hat{\theta}^A(s_G, x_L)(1-\mu_B).$$

From equations (5) and (6), we can obtain the simplification: $\mu_G = (1-\mu_B) = \theta p + \frac{1}{2}(1-\theta) > \frac{1}{2}$. The inequalities can then be verified, which proves that manager A's behavior is part of an equilibrium.[10]

Finally, it should be emphasized that herding equilibria exist in a model with any number of managers. Consider a third manager C, who moves after A and B, but before the state of the economy is realized. Since we have established that in a two-manager herding equilibrium, manager B's actions are independent of his signal, manager C learns nothing from observing what firm B has done. He learns only from observing firm A, whose actions do depend on the signal received. Thus manager C is in exactly the same position as manager B before him, and the same arguments can be used to show that he too cannot make his actions contingent on his private information. This line of reasoning can be applied repeatedly to any number of subsequent managers. Note, however, that other equilibria may also exist with more than two managers. For example, with three managers, the second may deviate from the first if he conjectures that the third manager will join him in a new herd.

III. Countervailing Forces

Up to this point, the model has been simplified to focus exclusively on the forces that make herd behavior likely. Naturally, there are other considerations which may offset herding tendencies. We discuss four such considerations below: managerial concern for profits; limited liability; wages that depend on relative, rather than absolute talent; and alternative definitions of ability.

The herding equilibrium derived above is generally inefficient relative to the first best for some configurations of private information. For example, if $x_H + x_L > 0$, then the equilibrium is inefficient when manager A observes s_B, manager B observes s_G, and firm B fails to invest. The expected profits lost by firm B due to herding, denoted by Π, are equal to $\frac{1}{2}(x_H + x_L)$.

If, however, managers place some weight on expected profits in their objective functions, these inefficiencies can disappear. Denote by R the amount by which the incentive constraint in (14) is violated. R measures the perceived reputational benefit to managers from herding, and it is proportional to

$$[\hat{\theta}(s_B, s_B, x_H) - \hat{\theta}(s_B, s_G, x_H)] + [\hat{\theta}(s_B, s_B, x_L) - \hat{\theta}(s_B, s_G, x_L)].$$

Managers who care only about their reputations will always herd, since $R > 0$. But managers who care about profits will have to trade off R against Π. For any given weight on profits in managers' objective functions, a large enough value of Π will restore the efficient signal-dependent equilibrium. Put differently, as the weight on profits increases, the range of parameter values over which there is herd behavior shrinks—the more egregious inefficiencies associated with herd behavior are alleviated, even though herding can still occur when Π is relatively small. This points to a role for short-term incentive contracts. Even if managers cannot be bound to their firms for life (so that they always place some weight on reputation) short-term contracts may shift the focus toward ex-

[10] It should be pointed out that with different parameter values, reputational concerns could induce even manager A to behave differently. Suppose that x_H is relatively large. Then efficiency could require manager A to invest after seeing s_G even if the posterior probability of success μ_G is small. Inspection of (21) reveals that this efficient equilibrium cannot be sustained if $\mu_G < \frac{1}{2}$; manager A will be unwilling to invest and thereby contradict the *ex ante* wisdom that success is very unlikely.

pected profits, thereby reducing herding tendencies.

Limited liability is another factor that can alter the tradeoff between reputation and profits. It may be that managers have fixed outside labor market opportunities and thus never have to accept a wage less than L. If L exceeds the wage to a manager with ability $\hat{\theta}(s_B, s_G, x_L)$ there is a floor on how poorly a manager can fare by disagreeing with his peers. Thus limited liability lowers R, the reputational gain from herding. Like profit-based compensation, limited liability shrinks the range of parameter values over which herd behavior is observed. A corollary is that herding may become more or less of a problem as a manager's career progresses. On the one hand, there is apt to be less uncertainty about the manager's ability, which should *reduce* the incentives for herd behavior. On the other hand, later in a successful career, wages are probably higher above the outside alternative L. This latter effect can *increase* the propensity to herd.

Relative ability concerns are a third factor that may offset herding tendencies. In our model, managers only care about having a high absolute ability assessment, $\hat{\theta}$. In certain situations, however, they may also care about how their $\hat{\theta}$ compares with that of other managers. For example, there may be a "superstars" effect present (see Sherwin Rosen, 1981), with top-ranking managers getting a disproportionately high wage. If this is the case, manager B will be more reluctant to mimic manager A, since by doing so he destroys any possibility of being the top-ranked manager.

A final way to attenuate the model's predictions regarding herd behavior would be to introduce a broader definition of ability. As we have defined it, ability is nothing more than an aptitude for making precise forecasts about the outcome of a *given* random variable. But ability might in addition include a knack for *uncovering* new random variables to study. Rather than having to decide simply whether or not to buy a certain asset, managers may also be responsible for finding alternatives to that asset. Under these circumstances, a desire to earn a reputation for "creativity" may deter managers from all choosing to buy the same asset.

IV. Implications of the Model

The theoretical model developed above has implications in a number of different areas. In order to give a feeling for some of the potential applications, we discuss a few examples.

A. *Corporate Investment*

Bank lending to LDCs was mentioned earlier as an apparent instance of herd behavior in corporate investment. A recent paper by Randall Morck, Andrei Shleifer, and Robert Vishny (1989) suggests that the potential for problems may be more widespread. They study the effectiveness of boards of directors in dealing with poorly managed firms. Their principal empirical finding is that top management firings are primarily associated with poor performance of a firm relative to its industry, rather than with industrywide failures. They interpret these results as evidence that boards have a difficult time assigning blame to their managers for mistaken strategies, when other firms in the industry are following similar strategies—at least in this segment of the labor market, there seems to be support for the "sharing-the-blame" effect discussed earlier. And as long as this effect is at work, one might expect that managerial behavior would be distorted in the direction of herding.

Herd behavior in corporate investment may have important consequences for the adoption of new technologies. Papers by Joseph Farrell and Garth Saloner (1985), and Michael Katz and Carl Shapiro (1985), have shown that compatibility externalities can lead to bandwagon effects in technology adoption. Our work can be viewed as complementary to theirs: even when compatibility is not an important concern, bandwagons can arise. If the manager of one firm adopts a particular technology, this creates a reputational externality, in the sense that other managers will tend to be biased toward the same technology for reputational reasons.

B. *The Stock Market*

Herd behavior by money managers could provide a partial explanation for excessive stock market volatility. By mimicking the behavior of others (i.e., buying when others are buying, and selling when others are selling) rather than responding to their private information, members of a herd will tend to amplify exogenous stock price shocks. In a sense, the ideas developed here can be thought of as providing the "microfoundations" for stock market phenomena that are often thought to stem from psychological sources such as "groupthink," mass euphoria, or panic.

It should be pointed out, however, that the model of this paper does not fit perfectly into a stock market setting, due to the assumption of perfectly elastic supply and the consequent lack of a market clearing price. Adding pricing considerations would complicate the formal analysis considerably. Nonetheless, we think that our basic insights do carry over to the stock market. At any given level of prices, money managers are likely to have an idea about the extent to which their competitors are "in" the market. If this is the case, there is the possibility that money managers will mimic each others' asset allocation strategies—upon observing that manager A has 50 percent of his assets in stocks and 50 percent in bonds, manager B may aim for a similar portfolio composition, even when his private information suggests that current price levels are too low or too high.[11] Thus one testable implication of our model is that the asset allocation decisions of professional money managers should be more closely correlated over time than the decisions of equally active private investors who are unconcerned about their reputations.

Robert Shiller and John Pound (1986) present some evidence that can be viewed as consistent with the existence of herd behavior in the stock market. They surveyed institutional investors to determine the factors that went into their decision to buy a particular stock. Purchase of stocks that had recently had large price run-ups tended to be motivated by the advice of others (other investment professionals, newsletters, etc.). This contrasted with more stable stocks, where fundamental research (a systematic search procedure for a security with certain characteristics) played a more important role. This suggests that the comfort inherent in following common wisdom can lead professional money managers to invest in stocks where fundamentals might dictate otherwise.

C. *Decision Making Within Firms*

Recent work on the theory of the firm by Raaj Kumar Sah and Joseph Stiglitz (1985, 1986) has emphasized the role of managers as information filters. They argue that firms may organize themselves internally in such a way as to take maximum advantage of the fact that different managers tend to have errors of judgment that are not perfectly correlated. Thus there may be benefits to having decisions made by committees, or through a vertical chain of command where projects can be rejected at various points along the way.

Our model points up certain limitations that may be inherent in group decision making, and also offers some new insights about how organizational structure can facilitate the decision-making process. As a stylized example, consider the case of a capital budgeting committee meeting, where the managers are supposed to vote in turn on a proposed investment project. Ideally, the point of having several managers vote is to gather a wide range of information. However, if career concerns are present, this may not work well. Once the first manager has voted, the others may simply echo his choice, regardless of their private beliefs. Thus a

[11] Another subtlety that is present in the stock market case but not in our model is a continuum of investment choices. A portfolio manager whose private information tells him that stocks are overpriced can "partially herd" by putting 45 percent, rather 50 percent of his assets in stocks. This may tend to dampen the aggregate effects of herd behavior.

false consensus is achieved, and the information of the other managers is wasted.[12]

One way around this problem is to have managers submit their votes simultaneously, perhaps in writing. However, this may limit valuable exchanges of ideas. An alternative approach is to have those with the stronger reputational concerns vote first. As we noted in the previous section, if the limited liability effect is not too important, reputational concerns will be strongest among young managers, since there is presumably more uncertainty about their ability. Thus if the committee consists of young and old executives, the young ones should be asked to voice their opinions before the old ones. More generally, this line of reasoning implies an advantage to a "bottom up," rather than "top down" organization of information flow within a firm. To the extent that new ideas can be passed upstream for approval, this may result in better decision making than if the ideas are originated at a high strategic planning level and then are passed downstream for line manager input.[13]

V. Conclusions

Herd behavior can arise in a variety of contexts, as a consequence of rational attempts by managers to enhance their reputations as decision makers. In addition to reputational concerns, there are other factors that influence herding. One of these is the extent to which there are commonly unpredictable components to investment outcomes: correlated prediction errors lead to the "sharing-the-blame" effect that drives managers to herd. Also important is the nature of the managerial labor market: herding is more likely to be a problem when managers' outside opportunities are relatively unattractive, and when compensation depends on absolute rather than relative ability assessment.

List of Notation

Symbol	Meaning
x_H	Investment proceeds in "high" state
x_L	Investment proceeds in "low" state
α	*Ex ante* probability of high state
s_G	Good signal
s_B	Bad signal
p	Prob($s_G \mid x_H$, smart)
q	Prob($s_G \mid x_L$, smart)
z	Prob($s_G \mid x_H$, dumb) = Prob($s_G \mid x_L$, dumb)
θ	*Ex ante* probability that manager is smart
$\hat{\theta}$	Posterior probability that manager is smart
μ_G	Prob($x_H \mid s_G$)
μ_B	Prob($x_H \mid s_B$)
Π	Profit passed up in herding equilibrium
R	Reputational benefit to herding
L	Wage that can be earned in outside opportunity

[12] There are other explanations for this "yes-man" effect. One is discussed in Solomon Asch's (1955) classic study of the effects of group pressure. He found that experimental subjects were reluctant to disagree with others in the group (confederates) despite the fact that the confederates' stated opinions were clearly wrong. Asch interprets the results as evidence in support of the view that individuals have an inherent psychological desire to conform to group norms.

[13] Keitaro Hasegawa (1986) discusses at length the emphasis placed by Japanese corporation on implementing a bottom-up flow of information. This form of organizational design is known there as "ringi."

REFERENCES

Asch, Solomon, "Opinions and Social Pressure," *Scientific American*, 1955, *193*.

Crawford, Vincent and Sobel, Joel, "Strategic Information Transmission," *Econometrica*, November 1982, *50*, 1431–51.

Farrell, Joseph and Saloner, Garth, "Standardization, Compatibility, and Innovation," *Rand Journal of Economics*, Spring 1985, *16*, 70–83.

Fudenberg, Drew and Tirole, Jean, "A Signal-Jamming Theory of Predation," *Rand Journal of Economics*, Autumn 1986, *17*, 366–76.

Gwynne, Samuel, *Selling Money*, New York: Weidenfeld and Nicolson, 1986.

Hart, Oliver and Holmstrom, Bengt, "The Theory of Contracts," in Truman Bewley, ed., *Advances in Economic Theory, Fifth World Congress*, New York: Cambridge University Press, 1987.

Hasegawa, Keitaro, *Japanese-Style Manage-*

ment: An Insider's Analysis, Tokyo: Kodansha, 1986.

Hellwig, Martin, "On the Aggregation of Information in Competitive Markets," *Journal of Economic Theory*, June 1980, *22*, 477–98.

Holmstrom, Bengt, (1982a) "Managerial Incentive Problems: A Dynamic Perspective," in *Essays in Economics and Management in Honor of Lars Wahlbeck*, Helsinki: Swedish School of Economics, 1982.

______, (1982b) "Moral Hazard in Teams," *Bell Journal of Economics*, Autumn 1982, *13*, 324–40.

______ **and Ricart i Costa, Joan,** "Managerial Incentives and Capital Management," *Quarterly Journal of Economics*, November 1986, *101*, 835–60.

Jensen, Michael, "Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers," *American Economic Review Papers and Proceedings*, May 1986, *76*, 323–29.

Katz, Michael and Shapiro, Carl, "Network Externalities, Competition, and Compatibility," *American Economic Review*, June 1985, *75*, 424–40.

Keynes, John Maynard, *The General Theory of Employment, Interest and Money*, London: Macmillan, 1936.

Milgrom, Paul and Stokey, Nancy, "Information, Trade, and Common Knowledge," *Journal of Economic Theory*, February 1982, *22*, 17–27.

Morck, Randall, Shleifer, Andrei and Vishny, Robert, "Alternative Mechanisms for Corporate Control," *American Economic Review*, December 1989, *79*, 842–52.

Nalebuff, Barry and Stiglitz, Joseph, "Information, Competition, and Markets." *American Economic Review Papers and Proceedings*, May 1983, *73*, 278–83.

O'Brien, Patricia, "Analysts' Forecasts as Earnings Expectations," *Journal of Accounting and Economics*, January 1988, *10*, 53–83.

Rosen, Sherwin, "The Economics of Superstars," *American Economic Review*, December 1981, *71*, 845–58.

Sah, Raaj Kumar and Stiglitz, Joseph, "Human Fallibility and Economic Organization," *American Economic Review Papers and Proceedings*, May 1985, *75*, 292–97.

______ **and** ______, "The Architecture of Economic Systems: Hierarchies and Polyarchies," *American Economic Review*, September 1986, *76*, 716–27.

Shiller, Robert and Pound, John, "Survey Evidence of Diffusing of Interest Among Institutional Investors," NBER Working Paper No. 1851, 1986.

Perfect Equilibria in a Trade Liberalization Game

By KIMINORI MATSUYAMA*

The credibility of temporary protection is examined in a simple infinite horizon, perfect information game of timing in which the domestic government uses the threat of future liberalization to induce the domestic firm to invest. All pure strategy subgame-perfect equilibria are cyclical and, surprisingly, one of them implements optimal temporary protection. However, this equilibrium fails to pass another credibility criterion called "Renegotiation-Proof." The game has a unique stationary subgame-perfect equilibrium in mixed strategies. (JEL 026, 422)

In trade policy debates, it is often argued that domestic industries should receive temporary protection from import competition. Immediate trade liberalization and ensuing inflows of foreign products and capital would jeopardize domestic firms, while protection would allow the domestic industries to introduce new technologies and products, thereby effectively competing with their foreign rivals. Any such protectionist measure should be temporary, because, under permanent protection, the lack of competitive pressure reduces incentives for domestic firms to rationalize their operations and to hold down costs. For example, this is the idea underlying the escape clauses in Article XIX of the General Agreement on Tariffs and Trade (GATT) and Sections 201 through 203 of the U.S. Trade Act of 1974.

A similar argument arises in the context of infant industry protection. Temporary support by governments sometimes helps new industries grow strong enough to meet international competition, but indefinitely imposed protection often results in perpetual industrial stagnation.

Despite its pervasiveness and (or perhaps because of its) persuasiveness, few studies have attempted to formalize this line of argument.[1] In a recent paper on infant industry protection (Matsuyama and Motoshige Itoh, 1986), we showed that it can be theoretically justified, using Michael Spence's (1979) model of a dynamic oligopoly market.

*Department of Economics, Northwestern University, 2003 Sheridan Road, Evanston, IL 60208. This is a substantially revised and shortened version of Matsuyama (1987). I would like to thank Dilip Abreu, Eric Maskin, Daniel Vincent, Michael Whinston, and participants of the Math Center Bag Lunch seminar at Northwestern University, the NBER Summer Institute, and the conference on "Designing Policies to Open Trade" at the University of California-Davis, for their insightful discussion and two anonymous referees for helpful comments. My special thanks go to one referee whose advice helps to improve exposition greatly. All remaining shortcomings are my own.

[1] In the standard literature of infant industry protection, the argument for temporary protection is based on the technological assumption that dynamic external economies, which could justify protection policies, disappear once the industry becomes mature. For example, Max Corden (1974, p. 256) discusses why protection should be temporary. "The temporary element can enter in three ways. (1) The learning may itself be temporary, being a characteristic of the firm's infancy period. (2) The imperfection of information or of the capital market, as these apply to the firm concerned, may be temporary: as the firm expands and its costs fall it may find it easier to finance further investments, whether in visible or invisible capital. (3) We may be constrained to the use of a tariff as a method of protection (the fiscal constraint ruling out direct or indirect export subsidization), so that the tariff could end once imports of the product have been completely replaced, and *should* end if the firm has monopoly power and above-normal profits are to be avoided." On the other hand, the popular argument in trade policy debates puts more emphasis on the incentive effect of temporary protection. For example, the study by the Organization for Economic Cooperation and Development (1985, p. 22) argues that "[p]rotection itself becomes less effective in promoting adjustment when—as a result of the repeated renewal of protectionist measures—the firms being protected have no reason to expect that they will even be exposed to the full challenge of international competition."

In this model, temporary protection works for two reasons. First, it provides the domestic firm with an opportunity to accumulate its capital stock. Second, the anticipation of future removal of the barrier (the temporary nature of protection) gives an incentive for the domestic firm to do so. We also showed that such a temporary protection policy can be optimal from the national welfare viewpoint.

However, we did not advocate adopting such temporary protection, because the effectiveness of temporary protection crucially rests on the presumption that the government can make a credible commitment on the future removal of protectionist measures, which we doubt is necessarily the case.[2] After all, there have been many protection policies, which were said to be temporary and turned out to be permanent. Indeed, this danger of prolonged protection is widely recognized. For example, Rachel McCulloch warns that:

> In principle, [temporary protection] policies provide breathing room for the affected industry, time in which to improve its competitive position or, in some cases, to phase out domestic production of goods where comparative advantage has shifted unambiguously abroad. In practice, however, neither adjustment to a status of full competitiveness nor phasing out of domestic production will necessarily occur during the limited period of import relief, so that the same industries return again and again for additional "temporary" relief. [1985, p. 154]

In this paper, I formalize this pitfall of optimal temporary protection: dynamic inconsistency.[3] To do so, I construct a simple infinite horizon, perfect information game of timing in which the domestic government and the domestic firm move alternately. The government prefers liberalizing the domestic market, given the firm's investment position. However, it is also willing to wait for another period if this could induce the firm to invest. The firm, if it believes that future liberalization is inevitable, prefers investing right before the liberalization in order to prepare for competition with its foreign rival. However, it may choose not to invest, hoping that this could induce the government to postpone the liberalization.

This game is meant to encompass a variety of situations. First, it could be the case of "infant" industries in developing countries whose governments decide whether to remove the inherited protection immediately or to wait until an announced future date. Second, it could be the case of "injured" industries in developed countries, where the governments consider granting temporary relief in each period. Third, it could be the case of "declining" industries whose protection is intended to ease pains of adjustment by allowing for gradual contraction of the industries. In the second and third cases, "liberalize" should be read as "refuse to grant import relief" and, in the third case, "invest" should be read as "reallocate." What is crucial in terms of formal analysis is that each player's action (but not its announcement) has irreversible consequences, which requires a game of timing framework, instead of a repeated game framework.[4]

[2] This is not the only reason why we did not advocate temporary protection. See the concluding section of Matsuyama and Itoh for other reasons.

[3] The seminal paper on the issue of dynamic inconsistency of optimal economic policies is Finn Kydland and Edward Prescott (1977). Most of recent developments center on monetary and fiscal policies. A few studies have addressed this issue in trade policies: see Maskin and David Newbery (1986), Robert Staiger and Guido Tabellini (1987) and Aaron Tornell (1987).

[4] In a simple game of timing, each player's only choice is when and whether to take a single pre-specified action, and the game ends once one player has moved. Two kinds of timing games are extensively discussed in the literature (Ken Hendricks and Charles Wilson, 1986, and Hendricks, Andrew Weiss, and Wilson, 1987). The first is the "war of attrition" developed in theoretical biology. Two animals fight over a territory. Fighting costs. Once one animal quits, its opponent gains the territory. This game has been applied to the problem of exit. See, for example, Pankaj Ghemawat and Barry Nalebuff (1985). In a war of attrition, each player prefers that the other moves first, but, if it has to move first, it prefers to do so sooner. On the other hand, in the second type of game of timing, "preemption game," each player prefers to move first,

The analysis presented below is partly motivated by the recent renewed interest in temporary protection as a way of diverting political pressure of protectionism: see, for example, Gary Hufbauer and Howard Rosen's (1986) and Robert Lawrence and Robert Litan's (1986) proposal of "refurbishing escape clauses." It is not at all clear what sort of political economic model they have in mind, but it appears quite difficult to deny any theoretical possibility of optimal temporary protection. The main theme of this paper is that adopting temporary protection may be a bad idea, *even if* it is optimal, due to the lack of credibility. And the central question asked is in what sense "optimal" policy lacks credibility. For this purpose, I intentionally do not describe situations in detail, which could give rise to an optimal policy of temporary protection. This is not to say that I believe temporary protection is generally optimal from the national welfare viewpoint. In fact, the government's objective is not necessarily identified as the national welfare below.

After presenting the game in Section I, the pure strategy Nash and subgame-perfect equilibria are discussed in Section II. Subgame perfection helps to eliminate all but a finite number of Nash equilibria, but, rather surprisingly, optimal temporary protection (and all Nash outcomes) can be supported by a subgame-perfect equilibrium, which suggests the inadequate power of the subgame-perfect restriction in this game. It is also shown that all subgame-perfect equilibria are cyclical.

In Section III, I turn to the stationary subgame-perfect equilibrium, which exists uniquely in mixed strategies. I will calculate the probability with which the optimal temporary protection succeeds and the expected length of the protection period, and show that the success rate is higher as the government is more impatient and as the firm is more patient and that, as both players are more patient, protection would last longer. The section ends by discussing an alternative interpretation of mixed strategies.

In Section IV, I employ another credibility criterion called "renegotiation-proof," recently developed by Joseph Farrell and Eric Maskin (1987) and David Pearce (1987). It will be shown that the subgame-perfect equilibrium that supports optimal temporary protection is not renegotiation-proof. Section V concludes the paper.

I. The Liberalization Game

Imagine the following scenario.[5] Initially, the domestic monopoly firm (player 2) earns its maximum profit in the protected domestic market. Then, a new government (player 1) takes office and the game starts. At the beginning of period 1, the government decides whether it liberalizes the market (L) or not (NL). If it chooses L, the foreign firm enters the market and both firms play the post-entry game from period 1 on. The liberalization game ends. If the government chooses NL, the domestic firm decides whether it invests (I) or not (NI). If it chooses I, the domestic firm earns less profit in period 1. The government will liberalize the market at the beginning of period 2 and the foreign firm will enter and both firms will play the post-entry game from period 2 on, with the domestic firm having the first-mover advantage. The liberalization game ends. If the domestic firm chooses NI,

but, it prefers to do so later. See Drew Fudenberg and Jean Tirole (1985). The two firms plan to introduce a new product to the market that can profitably support only one firm, and it would be cheaper to introduce the product later.

One can think of the game in this paper as a hybrid of a war of attrition and a preemption game; The government's payoff has the same property as in a war of attrition, while the firm's payoff has the same as in a preemption game. That is, the government prefers that the firm moves first (invest before liberalization), but if it has to move first (liberalize before investment), it prefers to do so sooner. On the other hand, the firm prefers to move first (invest before liberalization), but it prefers to do so later.

[5] This is only for fixing the idea. Other interpretations can easily be made, as suggested in the introduction.

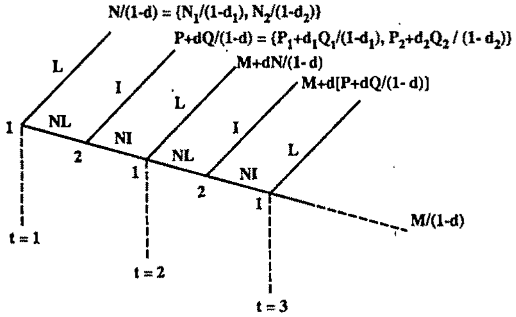

N/(1-d) = {N₁/(1-d₁), N₂/(1-d₂)}
P+dQ/(1-d) = {P₁+d₁Q₁/(1-d₁), P₂+d₂Q₂ / (1- d₂)}
M+dN/(1- d)
M+d[P+dQ/(1- d)]
L
I
L
I
L
NL
NI
NL
NI
1
2
1
2
1
M/(1-d)
t = 1
t = 2
t = 3

FIGURE 1. THE TRADE LIBERALIZATION GAME IN K. MATSUYAMA, "PERFECT EQUILIBRIA IN A TRADE LIBERALIZATION GAME"

then it earns its maximum profit in period 1. In this case, the government must decide between L and NL at the beginning of period 2 in the same situation as in period 1. This process continues until either the government chooses L or the domestic firm chooses I.

Figure 1 gives an extensive form representation of this game. Some explanations of notations are in order:

M: The government does not liberalize (NL) and the firm does not invest (NI). This is the status quo situation. Let M_1 and M_2 be the government's one-period payoff and the domestic firm's one-period profit in M, respectively.

N: The government liberalizes (L) before the firm invests. The firm plays the post-entry game with its foreign rival without the first-mover advantage. Let N_1 and N_2 be the government's and the domestic firm's one-period payoffs in N.

P: The government does not liberalize (NL) and the firm invests (I). Define P_1 and P_2 analogously.

Q: The government liberalizes (L) after the firm invests.. The firm plays the post-entry game with its foreign rival, with its first-mover advantage. Define Q_1 and Q_2 analogously.

Let $0 < d_1 < 1$ and $0 < d_2 < 1$ be the discount factors of the government and the firm. In Figure 1, payoffs of both players are given at each terminal node. For example, if the government chooses NL in period 1, the firm responds by NI, and the government chooses L in period 2, then the state of the market is M in period 1 and N from period 2 on. Thus, the payoff of the government is

$$M_1 + d_1N_1 + d_1^2N_1 + d_1^3N_1 + \cdots$$
$$= M_1 + d_1N_1/(1-d_1)$$

and the payoff of the firm is $M_2 + d_2N_2/(1-d_2)$.

To define the payoff functions more formally, let $g = \{g(1), g(2), \dots\}$ represent a strategy of the government, where $g(t)$ is either L or NL, its move in period t if the game has not ended. Likewise, let f be a strategy of the firm, with $f(t) = I$ or NI. (Only pure strategies are considered until Section III.) Define $m(g)$ as the smallest t satisfying $g(t) = L$ and $n(f)$ as the smallest t satisfying $f(t) = I$. (Let $m(g) = \infty$ if $g(t) = \text{NL}$ for all t and $n(f) = \infty$ if $f(t) = \text{NI}$ for all t.) Then, the payoff functions of the two players are given by

$$U_h(g,f) = \begin{cases} X_h(n(f)) & \text{if } n(f) < m(g), \\ Y_h(m(g)-1) & \text{if } m(g) \le n(f), \end{cases} \qquad (h=1,2)$$

where,

$$X_h(q) \equiv [1-d_h^{q-1}]M_h/(1-d_h) + d_h^{q-1}[P_h + d_hQ_h/(1-d_h)],$$
$$Y_h(q) \equiv [1-d_h^q]M_h/(1-d_h) + d_h^qN_h/(1-d_h)$$
$$= [1-d_h^{q-1}]M_h/(1-d_h) + d_h^{q-1}[M_h + d_hN_h/(1-d_h)].$$

Note that the outcome of the game, that is, the terminal node to be reached, depends solely on $m(g)$ if $m(g) \leq n(f)$ and solely on $n(f)$ if $m(g) > n(f)$.[6] When $m(g) = 1$, the outcome is *immediate liberalization* and the payoffs are given by $Y_h(0) = N_h/(1-d_h)$. When $m(g) = q+1 \leq n(f)$, it is an *unsuccessful q-period protection* and the payoffs are $Y_h(q)$. When $n(f) = q < m(g)$, it is a *successful q-period protection* and the payoffs are given by $X_h(q)$.

I make the following assumptions on the government's and the firm's one-period payoffs. First, given the investment position of the domestic firm, the government always prefers liberalization and the firm always prefers protection: (A1) $M_1 < N_1$, (A2) $P_1 < Q_1$, (A3) $M_2 > N_2$, and (A4) $P_2 > Q_2$. Note that (A2) is consistent with the assumption implicit above that, once the firm invests, the government liberalizes the market in the next period. Second, the firm has no incentive to invest without threat of future liberalization: (A5) $M_2 > P_2$. But, investment before liberalization gives the firm advantage in the post-entry game with the foreign firm: (A6) $Q_2 > N_2$. Note that assumptions (A4)–(A6) imply (A3). Finally, the government prefers the domestic firm having advantage over the foreign firm: (A7) $Q_1 > N_1$. Matsuyama and Itoh provide an example satisfying these assumptions, when the government's goal is identified as the national welfare.[7] Moreover, these assumptions appear consistent with the view of the world held by the advocates of escape clause protection: see Hufbauer and Rosen, and Lawrence and Litan.

I also assume that both players are sufficiently patient, since this game would be fairly trivial if either player discounts its future payoff heavily. For example, an impatient government might be unwilling to wait even for one period in order to induce the firm to invest. Then, the government can always reach its first best outcome by liberalizing immediately. To rule out this case, I assume

$$\text{(A8)} \quad Y_1(0) = N_1/(1-d_1) < P_1 + d_1 Q_1/(1-d_1) = X_1(1).$$

Given (A1), (A8) is also the condition for successful one-period protection to be the optimal outcome for the government.[8]

On the other hand, when the firm discounts its future payoff heavily, the government cannot induce the firm to invest. The threat of future liberalization would not work. Knowing this, the government would choose to liberalize the market in period 1. The following assumption helps to rule out this case.[9]

$$\text{(A9)} \quad X_2(1) = P_2 + d_2 Q_1/(1-d_2) > M_2 + d_2 N_2/(1-d_2) = Y_2(1).$$

From (A1), (A3), (A8), and (A9), $X_h(q)$ and $Y_h(q)$ can be shown to satisfy

$$\text{(P1)} \quad X_1(q) \text{ is decreasing in } q,$$

$$\text{(P2)} \quad X_2(q) \text{ and } Y_2(q) \text{ are increasing in } q,$$

$$\text{(P3)} \quad X_1(q) > Y_1(q-1),$$

$$\text{(P4)} \quad X_2(q) > Y_2(q),$$

$$\text{(P5)} \quad Y_1(0) > X_1(\infty).$$

[6] This does not mean that one can redefine the strategy sets so that a player's strategy is simply when it chooses to end the liberalization game. This is because the credibility of both the government and the firm's actions cannot be discussed without referring to the belief on what would occur in case one of the players fails to carry out its prescribed action.

[7] This example in Matsuyama and Itoh also satisfies that $P_1 > M_1$: the domestic firm under-invests. However, I do not use this assumption below.

[8] From (A2) and (A7), (A8) is equivalent to $(N_1 - P_1)/(Q_1 - P_1) < d_1 < 1$. Thus, the government might need to be patient to be interested in temporary protection. Note that N_1 may be smaller than P_1, in which case (A8) imposes no restriction. Matsuyama (1987, Proposition 1) shows that, when (A8) does not hold, any strategy with $g(1) = L$ is a dominant strategy and immediate liberalization is the dominant strategy outcome.

[9] From (A5)–(A6), (A9) is equivalent to $(M_2 - P_2)/[(M_2 - P_2) + (Q_2 - N_2)] < d_2 < 1$. Matsuyama (1987, Propositions 2 and 3) show that, when (A9) does not hold, immediate liberalization is the iterative dominant outcome.

With (P1)–(P5), we are now ready to characterize pure strategy equilibria.[10]

II. Subgame-Perfect Equilibria

Let us first look at Nash equilibria.[11] To do so, note that, from (P1), (P3), and (P5), there exists a unique positive integer q^* satisfying $X_1(q^*) \geq Y_1(0) > X_1(q^*+1)$. This is the maximal number of periods for which the government is willing to wait for the firm to invest. To simplify exposition, I assume away the nongeneric case $X_1(q^*) = Y_1(0)$.[12] Then,

$$\text{(P6)} \quad X_1(q^*) > Y_1(0) > X_1(q^*+1).$$

PROPOSITION 1: *The pure strategy Nash equilibria are characterized by either*

$$(1.1) \quad m(g)=1, \text{ and } n(f) \geq q^*+1,$$

or

$$(1.2) \quad 1 \leq n(f) = m(g)-1 \leq q^*,$$

which suggests that only immediate liberalization and successful q-period protection, where $1 \leq q \leq q^*$, *are pure strategy Nash outcomes.*

PROOF:

From the definition of q^*, the government's best response is $m(g)=1$ if $n(f) \geq q^*+1$; $m(g) > n(f)$ if $n(f) \leq q^*$. The firm's best response is $n(f) \geq m(g)$ if $m(g)=1$, since, when $m(g)=1$, its payoff does not depend on $n(f)$; $n(f) = m(g)-1$ if $m(g) \geq 2$, since $X_2(q)$ is increasing in q and $X_2(m(g)-1) > Y_2(m(g)-1)$ from (P4). Therefore, (1.1) and (1.2) are the only combinations of $m(g)$ and $n(f)$ which satisfy the mutual best response property. □

COROLLARY 1: *The optimal temporary protection can be supported by a Nash equilibrium with* $m(g)=2$ *and* $n(f)=1$.

PROOF:

It is immediate from setting $n(f)=1$ in (1.2).

At the equilibrium given in the corollary, the government is willing to wait in period 1 if the firm invests in period 1. The firm will invest in period 1 if it believes that the government will liberalize in period 2, *no matter what* its investment position is; the threat of liberalization works, if believed. Moreover, since the firm invests in period 1, the government incurs no cost by committing to the liberalization in period 2.

It is often argued, however, that the Nash equilibrium concept requires very weak rationality on the part of players, and therefore allows too many outcomes, some of which are unreasonable on intuitive grounds. In the context of the liberalization game, is it reasonable for the firm to believe that the liberalization in period 2 is inevitable? Is the government's commitment credible? If the government had to choose between L and NL in period 2 after the firm chose NI in period 1—which would never happen when the prescribed strategies $m(g)=2$, $n(f)=1$, are followed—does the government still have an incentive to stick to the prescribed move $g(2)=L$?

To examine whether the Nash equilibria possess this sort of credibility, I will look at subgame-perfect equilibria, following Reinhard Selten (1975).[13] Subgame perfection

[10] To characterize the equilibria, all I need is Assumptions (A1), (A3), (A8), and (A9). However, other assumptions are necessary to interpret (A8) and (A9) as the need for patience of the players.

[11] A pair of strategies (g, f) is called a *Nash equilibrium* if g maximizes the government's payoff given f and f maximizes the firm's payoff given g (the mutual best response property). A *Nash outcome* is a terminal node to be reached at a Nash equilibrium.

[12] See Matsuyama (1987) for the case of $X_1(q^*) = Y_1(1)$. It can be shown, from (A1), (A2), and (A7), that there exists a strictly increasing sequence $d(q)$ satisfying $d(1) = \text{Max}\{0, (N_1 - P_1)/(Q_1 - P_1)\}$, $d(\infty)=1$, and that $d(q^*) < d_1 < d(q^*+1)$ is equivalent to (P6). This implies that a more patient government is willing to wait longer. It also means that, for any large positive integer q, there exists a number $d \in (d(q), 1)$, such that the government whose discount factor is equal to d is willing to wait for more than q periods.

[13] A pair of strategies (g, f) is a *subgame-perfect equilibrium* of the game if its restriction to any subgame

means that no player can influence the outcome of the game by trying to make a threat which it would not carry out if called upon to do so. It turns out that subgame perfection helps to eliminate all but a finite number of Nash equilibria but it is not powerful enough to eliminate any of the Nash outcomes.

PROPOSITION 2: *There are q^*+1 pure strategy subgame-perfect equilibria, (g^k, f^k) $\{k=0,1,\ldots,q^*\}$, defined by,*

$$g^k(t) \equiv \begin{cases} L, & \textit{if } t \equiv k+1 \ (\operatorname{mod} q^*+1), \\ \mathrm{NL}, & \textit{otherwise}, \end{cases}$$

$$f^k(t) \equiv \begin{cases} I, & \textit{if } t \equiv k \ (\operatorname{mod} q^*+1), \\ \mathrm{NI}, & \textit{otherwise}, \end{cases}$$

which suggests that only immediate liberalization and successful q-period protection, where $1 \leq q \leq q^*$, are pure strategy subgame-perfect outcomes.

PROOF:

First, note that, for any subgame-perfect equilibrium (g, f), (P2) implies that $f(t) = I \rightarrow g(t+1) = L$ and (P4) implies that $g(t+1) = L \rightarrow f(t) = I$.

Now, suppose that $m(g) = 1$. From Proposition 1, $n(f) \geq q^*+1$ or $f(t) = \mathrm{NI}$ for $t = 1 \ldots q^*$. Then, from (P4), $g(t) = \mathrm{NL}$ for $t = 2 \ldots q^*+1$. This in turn implies $f(q^*+1) = I$ because, by applying Proposition 1 to the subgame starting at $t = 2$, $f(q^*+1) = \mathrm{NI}$ would imply $g(2) = L$, a contradiction. From (P2), $f(q^*+1) = I$ implies $g(q^*+2) = L$, which also guarantees that the restriction of (g, f) to the subgame starting at t $(3 \leq t \leq q^*+1)$ is a Nash equilibrium of the subgame. By repeating this process for the subgame starting at $t = 1+(q^*+1)z$ for any integer z, one can show that $g(t) = L$ if and only if $t = 1+(q^*+1)z$ and $f(t) = I$ if and only if $t = (q^*+1)z$. In other words, (g^0, f^0) is a subgame-perfect equilibrium and the only one that satisfies $m(g) = 1$.

One can show that (g^k, f^k) is the only subgame-perfect equilibrium that satisfies $m(g) = k+1$ and $n(f) = k$, for $k = 1 \ldots q^*$, in a similar manner. Finally, there is no Nash, and therefore, no subgame-perfect equilibrium with $m(g) \geq q^* = 2$, from Proposition 1. □

Proposition 2 shows that every pure strategy subgame-perfect equilibrium exhibits cycles with period q^*+1. Intuition behind the cyclicity should be clear. If the firm believes that the government will liberalize, it will invest. But, if the government believes that the firm will invest within q^* periods, it will wait. But, if the firm believes that the government will wait, it will not invest. But, if the firm will not invest within q^* periods, the government will not wait, *ad infinitum*. The proposition also states that there are q^*+1 subgame-perfect equilibria. This is due to the infinite, recursive nature of the game. Subgame perfection would not impose any restriction on the phase of the cycle.

Remark 1: Although the infinite nature of the game is crucial for the multiplicity of equilibria, the mechanism which generates the multiplicity in this game is different from that in infinitely repeated games. In an infinitely repeated game, the multiplicity arises only when each player's strategy is history-dependent. The equilibria shown in Proposition 2 are history-*independent*: neither player conditions its action on the past action by the opponent, which is payoff irrelevant. To see why the multiplicity arises despite history-independence, consider a T-period truncation of this game, Γ_T, where the players are barred from moving after period T. Then, there is the unique cyclical subgame-perfect equilibrium. (The government liberalizes in period T, which determines the phase of the cycle.) Obviously, the original game can be considered as Γ_∞, the limit of any subsequence of $\{\Gamma_T\}_{T=1}^{\infty}$. Let σ_T denote the unique equilibrium of Γ_T. From a theorem on limit games, such as Fudenberg and David Levine's (1983), the limit of equilibria

of the original game is a Nash equilibrium of the subgame. An outcome of the game is *subgame perfect* if there exists a subgame-perfect equilibrium that leads to this outcome.

of a sequence of games is an equilibrium of the limit of the games. The cyclicity implies that, although a sequence $\{\sigma_T\}_{T=1}^{\infty}$ does not converge, its subsequences, $\{\sigma_{k+1+(q^*+1)z}\}_{z=1}^{\infty}$ $\{k=0,1,2\dots q^*\}$ converge to different equilibria, (g^k, f^k) $(k=0,1,2\dots q^*)$, all of which are equilibria of the original game, Γ_{∞}, since it is the limits of $\{\Gamma_{k+1+(q^*+1)z}\}_{z=1}^{\infty}$ $\{k=0,1,2\dots q^*\}$. This shows that the cyclicity, not the history-dependence, is the source of multiplicity.

The next proposition is a direct corollary of Proposition 2.

PROPOSITION 3: *The optimal temporary protection can be supported by a subgame-perfect equilibrium* (g^1, f^1) *defined by* $g^1(t)=L$ *if and only if* $t\equiv 2 \pmod{q^*+1}$ *and* $f^1(t)=I$ *if and only if* $t\equiv 1 \pmod{q^*+1}$. *This is the only subgame-perfect equilibrium that supports optimal temporary protection.*

Proposition 3 states that liberalization in period 2 can be made credible through constructing a cyclical form of strategies after period 2. It also demonstrates that the only one of the Nash equilibria shown in the corollary of Proposition 1 is a subgame-perfect equilibrium.

The question is then: Are these equilibria "reasonable"? My answer is "Probably not." First, these equilibria have a "bootstrap" nature. The government changes its behavior periodically only because the firm does so, and the firm does so only because the government does so. Second, optimal temporary protection can be made credible only if immediate liberalization is credible. But, in the subgame-perfect equilibrium shown above, immediate liberalization can be made credible only because, if it would fail to liberalize, the government would punish itself by making a commitment not to liberalize for some time to come. This "self-punishing" property makes this equilibrium hardly convincing. Third, the cyclical nature of the strategies seems at odds with the following intuitive, although somewhat *ad hoc*, argument. If the firm does not invest in period 1, the situation the government faces in period 2 is exactly the same as in period 1, given the recursive structure of the game. If the government did not liberalize in period 1, how can one believe that it does not do the same in period 2? These considerations suggest a stronger restriction on equilibria: that is, each player must choose the same move when it faces the same situation. In the next section, I will turn to a stationary subgame-perfect equilibrium, which satisfies this requirement.

Finally, the equilibria in Proposition 2 are vulnerable to the possibility of renegotiation. I will turn to the problem of renegotiation-proofness in Section IV.

III. Stationary Subgame-Perfect Equilibrium

The stationarity requires that strategies should be time-independent. Once this property is imposed, one cannot hope that a subgame-perfect equilibrium exists in pure strategy space. In this section, I allow randomized strategies. Both players are assumed to be risk-neutral. Let $V=(V_1,V_2)$ represent the value of (the remainder of) the game for both players. In each round, the government chooses L with probability u and the firm chooses I with probability v. Let (u^*,v^*) denote equilibrium mixed strategies. From the argument above, $0<u^*<1$, $0<v^*<1$. This implies that both players are indifferent between their alternatives: that is,

$$(1)\quad P_2+d_2Q_2/(1-d_2)=M_2+d_2V_2$$

$$(2)\quad N_1/(1-d_1)=v^*\{P_1+d_1Q_1/(1-d_1)\}$$
$$+(1-v^*)(M_1+d_1V_1).$$

Furthermore, from the definition of $V=(V_1,V_2)$,

$$(3)\quad V_1=N_1/(1-d_1)$$

$$(4)\quad V_2=u^*N_2/(1-d_2)$$
$$+(1-u^*)$$
$$\times\{P_2+d_2Q_2/(1-d_2)\}.$$

Eliminating V_1 and V_2 from (1)–(4) gives.

$$(5)\quad N_1/(1-d_1) = v^*\{P_1+d_1Q_1/(1-d_1)\} + (1-v^*)\{M_1+d_1N_1/(1-d_1)\}$$

$$(6)\quad P_2+d_2Q_2/(1-d_2) = M_2+d_2[u^*N_2/(1-d_2) + (1-u^*) \times\{P_2+d_2Q_2/(1-d_2)\}].$$

There is a unique $0<u^*<1$ which satisfies (6) and a unique $0<v^*<1$ which satisfies (5). Furthermore, u^* is a function of d_2 but not of d_1, and v^* is a function of d_1 but not of d_2, which allows us to denote $u^*=u(d_2)$ and $v^*=v(d_1)$. One can also show that these functions have the following properties: $u(d_2)$ is decreasing in d_2, $u(d_2)\to 1$ as $d_2 \to \underline{d} \equiv (M_2-P_2)/[(M_2-P_2)+(Q_2-N_2)]$, and $u(d_2)\to 0$ as $d_2 \to 1$; $v(d_1)$ is decreasing in d_1, $v(d_1)\to 1$ as $d_1 \to d(1) \equiv (N_1-P_1)/(Q_1-P_1)$, and $v(d_1)\to 0$ as $d_1 \to 1$. In other words,

PROPOSITION 4: *There exists a unique mixed strategy stationary subgame-perfect equilibrium, in which the government chooses L with probability $u(d_2)$ and the firm chooses I with probability $v(d_1)$, where both u: $(\underline{d},1) \to (0,1)$ and $v:(d(1),1)\to(0,1)$ are decreasing, one-to-one functions.*[14]

It is easy to see why functions u and v have these properties. In order to make the choice between L and NL indifferent, a more impatient government (a low d_1) needs to be convinced that it is likelier that the firm would invest before liberalization, if another chance is given (a high v^*). Likewise, in order to be indifferent between I and NI, a more myopic firm (a lower d_2) needs to be convinced that future liberalization is likelier, even if it does not invest (a high u^*).

From (3), the value of the game to the government is $N_1/(1-d_1)$. It fails to achieve its maximum payoff due to its inability to make a credible commitment on the second-period liberalization:

$$(7)\quad V_1=Y_1(0)=N_1/(1-d_1) < P_1+d_1Q_1/(1-d_1)=X_1(1).$$

On the other hand, from (1) and (A9),

$$(8)\quad Y_2(0)=N_2/(1-d_2)<V_2 < P_2+d_2Q_2/(1-d_2)=X_2(1).$$

Thus, this equilibrium weakly Pareto-dominates immediate liberalization, while it is strictly Pareto-dominated by one-period successful protection.

The probability with which optimal temporary protection succeeds is

$$[1-u(d_2)]\,v(d_1).$$

The success rate is higher as the government is more impatient and as the firm is more patient. Another measure of "success" would be the probability with which the firm will eventually invest and the market ends up in Q. This is given by

$$(1-u^*)v^*/[1-(1-u^*)(1-v^*)],$$

which is also decreasing in d_1 and increasing in d_2. One can also calculate the expected length of protection. The market is liberalized in period 1 with probability u^*. The second-period liberalization has probability equal to $(1-u^*)v^*+(1-u^*)(1-v^*)u^* = (1-u^*)\{1-(1-u^*)(1-v^*)]$, In general, protection lasts for T periods with probability $(1-u^*)\{1-(1-u^*)(1-v^*)\}\{(1-u^*)(1-v^*)\}^{T-1}$. From this, the expected length of protection is

$$(1-u^*)/\{1-(1-u^*)(1-v^*)\},$$

which increases with d_1 and d_2. Therefore, if both players are more patient, protection would last longer.

To some economists, mixed strategies seem odd and less appealing, but they would appear more reasonable upon closer inspection. Certainly, it is hard to imagine the

[14] It remains an open question whether a nonstationary mixed strategy subgame-perfect equilibrium exists.

chairperson of the International Trade Commission (ITC) or the president casting dice to decide whether or not to protect the domestic industries not ready to face import competition. However, this aspect of the equilibrium is attributable to the fact that there is no chance move in the model. To see this, note that randomizing matters, not because it would affect the distribution of the player's payoff directly, but because it would keep the opponent uncertain about the player's choice.[15] Therefore, a random disturbance to the player's payoff, which is unknown when the opponent moves but known when the player moves would do.[16] For example, suppose that, in each period, uncertainty about the domestic demand resolves after the government's move and before the firm's. The firm can make a choice deterministically contingent on the market condition, but the government knows only the probability distribution of the firm's choice. Likewise, the government's decision in the next period would reflect new developments on economic and political situations after the firm's investment decision in this period. Then, even if the government uses pure strategies, the firm is uncertain about the government's move.

In reality, it is highly uncertain whether a domestic industry would receive temporary protection. For example, Lawrence and Litan (1986, Table 3-2) reports the U.S. experience of escape clause relief. It shows that, between 1975 and 1985, the presidents have granted domestic industries import relief in only fourteen of the 33 cases which the ITC affirmed import injury. It is not hard to imagine that the presidents made their decisions taking into account many factors which did not exist when these industries had brought the cases before the ITC.

IV. Renegotiation-Proof

This section examines the credibility of optimal temporary protection by invoking "renegotiation-proof," the concepts recently proposed by, among others, Farrell and Maskin, and Pearce. Although they developed their theories in the context of infinitely repeated games, one can easily apply these concepts to the liberalization game due to its recursive structure.

A noncooperative solution like a subgame-perfect equilibrium is sometimes interpreted as a "self-enforcing agreement." Imagine the situation where players can freely discuss their strategies, but cannot make binding commitments. Then any viable agreement must be self-enforcing: that is, once players agree to play equilibrium strategies, no player has an incentive to renege. In games with many equilibria as the liberalization game, this interpretation is appealing as a way of explaining how players know which equilibrium is to be played. Moreover, in the liberalization game, if the government and the firm were barred from communicating with each other, they could hardly coordinate their strategy choices so that one of the cyclical equilibria given in Proposition 2 would emerge. However, this ability to communicate itself would make these equilibria untenable. To see why, consider a subgame-perfect equilibrium supporting immediate liberalization. From (P1), (P2), (P4), and (P6),

$$(9)\quad X_1(q) > Y_1(0),$$
$$X_2(q) > Y_2(0) \quad \text{for all } 1 \leq q \leq q^*;$$

that is, immediate liberalization is Pareto-dominated by other subgame-perfect outcomes. Therefore, it is hard to imagine that the government and the firm agree upon this outcome. But, once they recognize this, other subgame-perfect equilibria would become also untenable. More specifically, consider optimal temporary protection in Proposition 3. If the firm reneges and does not invest in period 1, the government is supposed to liberalize in period 2. But, at the beginning of period 2, the firm could make a proposal

[15] This also explains why the probability distribution of the player's choice depends on the opponent's discount rate, but not on the player's.

[16] John Harsanyi (1973) proves that, in a static game context, a mixed strategy equilibrium in a game with deterministic payoffs can be approximated as a pure strategy equilibrium in a game with randomly disturbed payoffs.

of moving to an alternative subgame-perfect outcome which is more attractive to both the government and the firm than immediate liberalization. If the firm realized that it could renegotiate with the government in this way, it might find not to invest advantageous. Therefore, optimal temporary protection equilibria would collapse. Likewise, all equilibria in Proposition 2 have this problem. Farrell and Maskin propose a credibility criterion to rule out this sort of equilibria.[17]

In objecting to this one might argue that the firm cannot effectively negotiate with the government when it failed to invest, for the following reason. For example, consider the case $q^* = 1$. There are only two pure strategy subgame-perfect outcomes: immediate liberalization and optimal protection. To ask the government to wait for one more period, the firm needs to argue that optimal protection is a credible outcome. But this requires that immediate liberalization also be credible, since the credibility of optimal protection depends on that of immediate liberalization. In other words, in order to renegotiate away from immediate liberalization, the firm has to argue that it is *not* possible to renegotiate away from it.

However, the firm can negotiate to abandon the immediate liberalization equilibrium in favor of the stationary subgame-perfect equilibrium discussed in Section III. From (7) and (8), this equilibrium (weakly) Pareto dominates immediate liberalization and its credibility does not rely on that of immediate liberalization. If the firm can renegotiate away from liberalization this way, then all equilibria in Proposition 2, and therefore optimal protection, are not credible. This is in the spirit of Pearce's concept of renegotiation-proof.

In summary, optimal temporary protection is not renegotiation-proof in the sense of Farrell and Maskin. It is not renegotiation-proof in the sense of Pearce when mixed strategies are allowed for.

V. Concluding Remarks

In this paper, I have formalized dynamic inconsistency of optimal temporary protection, by constructing a simple, infinite horizon, perfect information game of timing in which the domestic government and the domestic firm move alternately. I examined subgame-perfect equilibria under the assumptions which guarantee that optimal temporary protection can be supported by a Nash equilibrium. Subgame perfection helps to eliminate all but a finite number of equilibria, but, rather surprisingly, optimal temporary protection can be supported by a subgame-perfect equilibrium. This is because the threat of period 2 liberalization can be made credible through constructing a particular cyclical form of strategies after period 2. I do not find this equilibrium compelling. One reason is that it is not renegotiation-proof. There exists a unique stationary subgame-perfect equilibrium in mixed strategies. In this equilibrium, the government's payoff is smaller than in the first best outcome, due to its inability to make a credible commitment.

However, I do not mean to say that the government cannot make a credible commitment to the future liberalization in reality. There are several possibilities which the liberalization game fails to take into account. First, the government might be aware of "the demonstration effect" of liberalization. If infinitely many industries ask for import relief sequentially, the government can and has incentive to build a reputation to be tough through a similar mechanism as the Folk Theorem in the repeated Prisoner's Dilemma game. Such a reputational mechanism can be designed in a renegotiation-proof way. Second, the government might be able to sign a contract with a third party (perhaps, the GATT) to make the cost of postponement of liberalization (or renewal of temporary protection) prohibitively high. Third, the domestic government might want to ask a foreign government to exert diplomatic pressure to liberalize the domestic market (as some observers suspect that Japanese Ministry of International Trade and Industry (MITI) has done with the United

[17] In the terminology of Farrell and Maskin, all pure strategy equilibria of this game are not "weakly renegotiation-proof."

States). Of course, these mechanisms should be analyzed more formally by explicitly modifying the game between the government and the firm, which I will leave on the agenda for future research.

Throughout the paper, I have asked whether a particular equilibrium is reasonable by investigating its credibility. An alternative criterion of "reasonableness" would be its robustness. If a set of possible equilibria disappears when the game is slightly perturbed in a natural way, one cannot rely on such equilibria in prediction. For one can never be sure of the exact nature of the games governments and firms play in reality. Thus, it is highly desirable to examine the robustness of equilibria by, say, introducing informational asymmetry into the game. For example, one can consider the case where, with some probability, the firm's investment effort might fail to materialize and the government cannot observe the firm's action.

Another natural extension would be to allow the firm to accumulate capital gradually. However, such a modification would make the model intractable, since one cannot invoke recursivity when each stage game depends on the level of capital stock. One way of getting around this difficulty is to assume that the number of periods is finite: see Tornell (1987). I find, however, that the finite period assumption is at least as unrealistic as the binary restriction of the firm's action space.

Perhaps the most problematic feature of the model is its treatment of the domestic government as a unified, coherent body of decision makers. In reality, any economic policy is a product of complex interactions among different parts of the government, each of which has its own objectives. In fact, a number of recent studies deal with the political economy of protectionism, as exemplified by Robert Baldwin (1985). The approach adopted here should be regarded as a complement of, not a substitute for, this growing body of literature.

Finally, some mention should be made regarding two other possible solutions to avoid permanent protection. First, some observers suggest that temporary protection should be granted to domestic industries only when they commit to adjustment plans. This view is reflected in a number of recent congressional proposals to reform the U.S. escape clause. In the context of the liberalization game, this amounts to enforcing the domestic firm to invest during the protection period. This approach, however, requires the government to participate actively in the retooling process of the industries. The danger of such government intervention in practice might be substantial and it is well documented.[18] Second, one might hope that the danger of prolonged protection can be reduced by designing appropriate institutional arrangements. In fact, it is often argued that the escape clause itself is "a useful safety valve for protectionist pressures (Kenneth Dam, 1970, p. 106)." However, this approach is, at best, an imperfect solution to the problem. As long as temptations to provide further protection remain, there exists a tendency to circumvent the institutional arrangements, as suggested by the recent drift of U.S. practice away from the escape clauses in negotiating orderly market arrangements and voluntary export restraints.

[18] See U.S. Council of Economic Advisors (1984, pp. 108–9) and Lawrence and Litan (1986, pp. 84–96).

REFERENCES

Baldwin, Robert E., *The Political Economy of U.S. Import Policy*, Cambridge, MA: MIT Press, 1985.

Corden, W. Max, *Trade and Economic Welfare*, Oxford: Oxford University Press, 1974.

Dam, Kenneth W., *The GATT: Law and International Economic Organization*, Chicago: University of Chicago Press, 1970.

Farrell, Joseph and Maskin, Eric, "Renegotiation in Repeated Games," Harvard University Discussion Paper No. 1335, 1987.

Fudenberg, Drew and Levine, David, "Subgame-Perfect Equilibria of Finite and Infinite Horizon Games," *Journal of Economic Theory*, December 1983, *31*, 251–68.

______ **and Tirole, Jean,** "Preemption and Rent Equalization in the Adoption of a

New Technology," *Review of Economic Studies*, July 1985, *52*, 383–401.

Ghemawat, Pankaj and Nalebuff, Barry, "Exit," *Rand Journal of Economics*, Summer 1985, *16*, 184–94.

Harsanyi, John, "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points," *International Journal of Game Theory*, 1, 1973, *2*, 1–23.

Hendricks, Ken, Weiss, Andrew and Wilson, Charles, "The War of Attrition in Continuous Time with Complete Information," Hoover Institution Working Paper No. E-87-50, 1987.

______ **and Wilson, Charles,** "Equilibrium in Preemption Games with Complete Information," Hoover Institution Working Paper No. E-86-72, 1986.

Hufbauer, Gary Clyde and Rosen, Howard F., *Trade Policy for Troubled Industries*, Washington: Institute for International Economics, 1986.

Kydland, Finn E. and Prescott, Edward C., "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

Lawrence, Robert Z. and Litan, Robert E., *Saving Free Trade: A Pragmatic Approach*, Washington: The Brookings Institution, 1986.

Maskin, Eric and Newbery, David, "Disadvantageous Oil Tariffs and Dynamic Consistency," mimeo., Harvard University, 1986.

Matsuyama, Kiminori, "Perfect Equilibria in a Trade Liberalization Game," Center for Mathematical Studies in Economics and Management Science Discussion Paper No. 738, Northwestern University, 1987.

______ **and Itoh, Motoshige,** "Protection Policy in a Dynamic Oligopoly Market," mimeo., Harvard University, 1986.

McCulloch, Rachel, "Trade Deficits, Industrial Competitiveness, and the Japanese," *California Management Review*, Winter 1985, *27*, 140–56.

Pearce, David, "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," mimeo., Yale University, 1987.

Selten, Reinhard, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1, 1975, *4*, 25–55.

Spence, A. Michael, "Investment Strategy and Growth in a New Market," *Bell Journal of Economics*, Spring 1979, *10*, 1–19.

Staiger, Robert and Tabellini, Guidos, "Discretionary Trade Policy and Excessive Protection," *American Economic Review*, December 1987, *77*, 823–37.

Tornell, Aaron, "Time Inconsistency of Protectionist Programs: On the Ineffectiveness of Investment-Contingent Subsidies," mimeo., Columbia University, 1987.

Organization for Economic Cooperation and Development, *Costs and Benefits of Protection*, Paris: OECD, 1985.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington: USGPO, 1984.

Differential Payments Within a Bidder Coalition and the Shapley Value

By DANIEL A. GRAHAM, ROBERT C. MARSHALL,
AND JEAN-FRANCOIS RICHARD*

Bidder coalitions at English auctions frequently distribute collusive gains among members via a secondary auction or "knockout." When coalition members are sufficiently heterogeneous, nested coalition structures are observed in which a knockout is conducted at each level of nesting. The nested knockout's characteristics are investigated. Within many settings we find that the expected payments to coalition members via the nested knockout equal the Shapley value. Incentive compatibility problems of the nested knockout are also analyzed.

Our understanding of auction schemes has progressed significantly since William Vickrey's (1961) seminal work. Recent research has focused on the optimal design of auctions as well as the strategic behavior of bidders and auctioneers within specific environments. A rich set of results has emerged from these investigations. However, one frequently observed characteristic of strategic behavior remains relatively unexplored to date—bidder collusion.

Within the independent private values model (IPV) it is well known that the dominant noncooperative strategy for each bidder at an oral ascending bid (English) auction is to remain active until the bid reaches his or her personal valuation for the item being auctioned. Similarly, at a second price auction it is a dominant noncooperative strategy for each bidder to submit a bid equal to his personal valuation. Restricting attention to the English and second price auctions, it is also well known that outcomes attainable through collusive behavior strictly dominate those associated with noncooperative play. Intuitively, a bidder coalition or ring, which consists of k of the n bidders attending an auction, can expect to gain, relative to noncooperative behavior, by suppressing $k-1$ of the competitive bids at the main auction. Daniel A. Graham and Robert C. Marshall (1987), hereafter GM, show that when the members of a bidder coalition are homogeneous it is possible for them to organize themselves in such a way that participation is mutually advantageous and yet there is no advantage to cheating. These facts correspond well to many of the "stylized facts" of ring behavior summarized in Section I. In practice, rings allocate items won at a main auction through a secondary auction called a knockout. If member types are considered to be relatively homogeneous, then only a single knockout is observed. In this situation the difference between the price paid for the item at the knockout and the price paid for the item by the coalition at the main auction is *equally* divided among all coalition participants. Within the IPV model where bidders draw their valuations from the same distribution, GM have proposed a mechanism through which the gains from coalition formation are equally distributed among all ring members.

In practice, when ring members are heterogeneous, it is common to observe a nested coalition structure. This structure facilitates differential payments to ring members.

*Professor of Economics, Associate Professor of Economics and Professor of Statistics and Decision Sciences, Duke University, Durham, NC 27706. The authors are indebted to Jim Friedman, Roger B. Myerson, and Alvin Roth for many helpful comments. Financial support from the National Science Foundation, grant no. SES-8708615, is gratefully acknowledged. The last two authors also acknowledge support from the Pew Charitable Trust and the Ford Foundation.

Specifically, a knockout is conducted at each level of nesting—the details of the procedure are described in Section I.

The goal of this research is to investigate the characteristics of the nested knockout and the factors underlying its use by practicing bidder coalitions. Attention is devoted to a single-object English (or second price) auction. To facilitate an initial understanding of the nested knockout, in Section II we assume that the valuations of bidders are commonly known by all bidders and that the levels of nesting equal the number of bidders within a ring. Representing this complete information game in characteristic function form yields a remarkable result—the nested knockout produces an imputation that not only belongs to the core of the cooperative game but also is identical to the Shapley value of the game. This particular distribution of ring benefits is also explained in terms of Roger B. Myerson's (1980) "cooperation structures" as the result of cooperative agreements in which the parties to the agreement share equally the benefits derived from the agreement.

In Section III we not only adopt a more realistic setting in which valuations are private information but we assume, as observed in practice, that the knockout is conducted after the main auction (i.e., an *ex post* knockout) and without the use of the budget breaking ring "center" employed by GM. To simplify the analysis it is assumed that the ring members draw their valuations from the same distribution or, in other words, that they are *ex ante* homogeneous. The knockout is modeled as one in which the low bid is divided among all members, the difference between the second lowest and the lowest is divided among all members save the one who submitted the lowest bid, and so forth. Obvious incentive compatibility problems arise in this context. Specifically, low-valued bidders have an incentive to bid in excess of their true valuations at the post-auction knockout in hopes of enlarging their share of the "pie." We investigate the stategic behavior of ring participants when side payments are determined by the nested knockout in both a simultaneous and sequential game framework. Surprisingly, in both the simultaneous and sequential versions of the nested knockout, a ring member with valuation v receives equal *ex ante* expected payoffs (for any value of v). This result thus extends the GM "equal shares" result to mechanisms that are far more descriptive of the nested knockouts used in practice.

Some initial results are presented in Section IV regarding the unequal sharing of expected ring benefits among *ex ante* heterogenous bidders, that is, bidders who draw their valuations from different, but commonly known, distributions. These ring members adopt an incentive efficient and durable mechanism, in the sense of Bengt Holmström and Roger B. Myerson (1983), in which payments differ among the ring members—the payment to a particular ring member depends upon the distribution from which he draws his valuation. It is demonstrated that these payments yield an imputation that is once again identical to the Shapley value of the characteristic function game based upon expectations of payoffs, provided that cooperation by ring members entails agreements to share equally the gains resulting from their cooperation. This result is also related to that obtained by Myerson (1980).

I. The Stylized Facts of Bidder Collusion

The character of a bidder coalition depends upon the type of object being sold. Ralph Cassady, Jr. (1967, ch. 13), discusses the many types of rings found throughout the world. Below we provide the stylized facts of cooperative behavior at an English auction where a nondivisible item is being sold. This account is not conjecture but is based, in fact, upon information provided by both auctioneers and ring members. For further details we refer the interested reader to Graham and Marshall (1985). GM explains Facts 1–6 within the context of a single object English auction where bidders' valuations are independently and privately drawn from the same distribution. The present research focuses on Fact 7.

1. Rings exist and have a stable form of organization over time.
2. Rings adopt strategies that eliminate meaningful competition among members

at the main auction and yet ensure that no item will be sold to a non-ring bidder or be retained by the auctioneer at a price below the maximum of the individual ring members' personal valuations.

3. Rings have open membership policies in the sense that bidders who are expected to be competitive at the main auction are invited to join.
4. The auctioneer responds strategically to the existence of a ring.
5. Rings attempt to conceal their existence from the auctioneer.
6. The benefits of ring formation are shared among the members rather than, for example, accruing entirely to the ring member who ultimately obtains possession of the item.
7. Since the expected benefits to the ring of a particular bidder's membership often varies across bidders, the ring frequently adopts a procedure by which the benefits of ring formation are shared unequally among its members. A nested coalition structure is frequently observed in which the individuals expected to make the greatest contribution, *ex ante*, are typically found in the innermost ring, and those expected to make the smallest contribution are found in the outermost ring.

The details of the nested knockout are best described with an example. Suppose that a coalition of ten bidders has formed and appointed a sole bidder to bid on their behalf at the main auction. Suppose further that the sole bidder has won the item at the main auction for p_0. These ten bidders become the members of an initial or outer ring. Four levels of nesting develop: within the initial ring with ten members there is a ring of size eight, within the ring of size eight there is a ring of size six, and within the ring of size six there is a ring of size four. Ignore for the moment the decisions of individuals to participate in these rings—this will be addressed in Sections II and III—and call these rings R_1, R_2, R_3, and R_4, respectively. Note that the rings are *nested* in the sense that $R_{i+1} \subset R_i$. Finally, let $R_i \backslash R_{i+1}$ denote those bidders who are members of ring R_i but not members of ring R_{i+1}.

The nested coalition structure makes possible a series of knockouts in which English auctions are used to determine the ultimate ownership of the item. In the first knockout the members of R_2 bid collusively against those in $R_1 \backslash R_2$. In practice the members of R_2 would appoint a sole bidder who would remain active in the bidding up to the highest valuation within the members of R_2. The members of $R_1 \backslash R_2$ would bid competitively. Let p_1 denote the winning bid. The individual (or sole bidder) who bid p_1 wins the item in the first knockout and pays p_1 for it. The "surplus" $p_1 - p_0$ is then equally divided among all members of R_1. (p_0 is paid to the sole bidder of the original coalition who paid this amount to the auctioneer at the main auction for the item.) If the winning bid came from a member of $R_1 \backslash R_2$, then the knockouts are finished. On the other hand, if p_1 were submitted by the sole bidder of R_2, then the item becomes the property of R_2 and another knockout takes place. In the second knockout only members of R_2 participate—bidders in $R_1 \backslash R_2$ have left with their side payments by the time this knockout occurs. Now the sole bidder for R_3 will bid against the members of $R_2 \backslash R_3$. If p_2 denotes the winning bid then $p_2 - p_1$ is equally divided among all members of R_2. If p_2 came from a member of R_2 / R_3, then all is done—if not, the item becomes the property of R_3 and another knockout will be conducted. It is possible for the knockouts and side payments to continue in this manner until the members of R_4 have acquired the item. Should this happen, the members of R_4 would then conduct the final knockout, bidding competitively among themselves to determine ultimate individual ownership of the item.

Table 1 illustrates the side payments produced by the nested knockout for the case in which the item reaches the final knockout by R_4. These payments sum to $p_4 - p_0$. In addition, the ultimate winner of the item receives the usual surplus: the winner's valuation, $\bar{v}$, minus p_4. Consequently, the winner in ring 4 receives

$$\text{Ring 4 Winner: } x_4^W = x_4 + (\bar{v} - p_4).$$

This illustrates the nested knockout. A general definition will be provided in Section II.

TABLE 1—SIDE PAYMENTS IN THE NESTED KNOCKOUTS

Knockout	$R_1 \backslash R_2$	$R_2 \backslash R_3$	$R_3 \backslash R_4$	R_4
1st	$\frac{p_1 - p_0}{10}$	$\frac{p_1 - p_0}{10}$	$\frac{p_1 - p_0}{10}$	$\frac{p_1 - p_0}{10}$
2nd		$\frac{p_2 - p_1}{8}$	$\frac{p_2 - p_1}{8}$	$\frac{p_2 - p_1}{8}$
3rd			$\frac{p_3 - p_2}{6}$	$\frac{p_3 - p_2}{6}$
4th				$\frac{p_4 - p_3}{4}$
Total	$x_1 = \frac{p_1 - p_0}{10}$	$x_2 = x_1 + \frac{p_2 - p_1}{8}$	$x_3 = x_2 + \frac{p_3 - p_2}{6}$	$x_4 = x_3 + \frac{p_4 - p_3}{4}$

The use of the nested knockout by practicing bidder coalitions leads to a number of questions. Does the nested knockout possess unique properties that recommend its use relative to other surplus sharing mechanisms? What type of bidding behavior would be induced by a nested knockout given that high losing bids are advantageous to low-valued ring members? If incentive compatibility cannot be obtained within the nested knockout as described above, then is it possible to construct an incentive compatible and efficient mechanism for the allocation of the collusive gain? In the following sections these questions are addressed. First, in Section II we characterize the payments of the nested knockout in an environment that allows us to suppress incentive compatibility concerns. In Section III we move to an IPV framework and investigate bidding behavior for two variants of the nested knockout. Finally, in Section IV we pose a variant of the nested knockout where payments are made *ex ante* by a ring "center" instead of being determined *ex post* via ring members' bids. This mechanism is both incentive efficient and durable.

II. The Complete Information Auction Game

We initially consider a highly simplified knockout game in which all of the ring members at an English auction have valuations that are known to one another. This "public values" model is extremely unrealistic but (i) illustrates an important feature of the nested knockout procedure and (ii) will be relaxed in the next section to allow for valuations to be private information. For notational convenience we assume that the n_0 initial ring members are ordered by their valuations so that

$$(1) \qquad v_1 \geq v_2 \geq \cdots \geq v_{n_0}.$$

Suppose that the ring has acquired the item at a price equal to p_0. Given the complete information of members valuations we may suppose that $p_0 \leq v_1$ and let n denote the number of ring members whose valuations exceed p_0:

$$n \equiv \max\{k | v_k \geq p_0\}.$$

These n bidders will be the only *active* participants of the ring in the complete information auction game. The nested knockout imputation for this game is easily defined. Since valuations are public knowledge, perfectly fine nesting is possible. By perfectly fine we mean that ring R_1 will contain $\{1,\dots,n\}$, R_2 will contain $\{1,\dots,n-1\}$, etcetera. The kth knockout then pits the members of $R_{k+1} = \{1,\dots,n-k\}$ against $R_k \backslash R_{k+1} = \{k\}$. The imputation resulting from the nested knockout procedure gives bidder i

$$x_i = \sum_{j=i}^{n-1} \frac{v_j - v_{j+1}}{j} + \frac{v_n - p_0}{n} \qquad \text{for } i < n,$$

$$x_n = \frac{v_n - p_0}{n},$$

so that

$$x_i = x_{i+1} + \frac{v_i - v_{i+1}}{i} \quad \text{for } i < n.$$

Note that the conventional specification of the characteristic function would assign the maximin value to a coalition S in the game played between S and the complementary coalition $N-S$. However, since the maximin value is zero for any ring that is not all-inclusive this approach would yield the rather uninteresting characteristic function that assigns zero to all coalitions other than N. Our specification of the characteristic function, on the other hand, is based upon the fact that it is a dominant strategy in the game between S and $N-S$ for S to remain active until the bidding reaches

$$\max_{i \in S} v_i$$

and for $N-S$ to remain active until the bidding reaches

$$\max_{j \notin S} v_j.$$

Since the payoff to S under these dominant strategies is

$$(2) \quad w(S) \equiv \max\left\{\max_{i \in S} v_i - \max_{j \notin S} v_j, 0\right\} \quad \forall S \subseteq N,$$

we take this to be the relevant characteristic function for the complete information game. Given the indexing convention (1), the characteristic function can also be expressed as

$$(3) \quad w(S) = \begin{cases} 0 & \text{if } 1 \notin S \\ v_1 - v_i & \text{if } 1,2,\ldots,i-1 \in S \\ & \text{and } i \notin S \end{cases}$$

Recall that an imputation, X, is in the core of w iff

$$\sum_{i \in S} x_i \geq w(S) \qquad \forall S \subseteq N.$$

It follows that

THEOREM 1: *The nested knockout imputation is in the core of the complete information auction game.*

PROOF:

See Appendix, Part A.

The nested knockout imputation, which gives larger payments to bidders with larger valuations, belongs to the core. An equal division of $w(N)$ among the members of the ring, on the other hand, does not generally belong to the core. Suppose, for example that $n = 3$, the valuations are

$$(v_1, v_2, v_3) = (12, 7, 3)$$

and $p_0 = 0$. Then the characteristic function for this game is

$$w(\{1\}) = 5 \qquad w(\{2\}) = 0$$
$$w(\{1,2\}) = 9 \qquad w(\{1,3\}) = 5$$
$$w(\{1,2,3\}) = 12$$
$$w(\{3\}) = 0$$
$$w(\{2,3\}) = 0$$

and the nested knockout imputation is

$$(x_1, x_2, x_3) = (8, 3, 1).$$

An equal division of $w(\{1,2,3\})$ would give each bidder a payoff of 4. That this is not in the core follows from the fact that $w(\{1,2\}) = 9 > 4+4$. On the other hand, the "winner take all" imputation, (12,0,0) in this example, is in the core.[1] The nested knockout imputation has additional properties which distinguish it from this and other core imputations.

[1] The set of imputations in the core is defined by the inequalities $x_1 \geq 5$, $x_2 \geq 0$ and $12 \geq x_1 + x_2 \geq 9$ together with the identity $x_1 + x_2 + x_3 = 12$. It follows, in particular, that we can have $x_2 \geq x_1$ or $x_3 \geq x_2$ but not both. When the valuations are not too far apart, the core includes "decreasing" imputations, for example, if $(v_1, v_2, v_3) = (12, 9, 7)$ then the imputation (3.5,4,4.5) is in the core.

It is also useful to compare the nested knockout imputation to the Shapley value for this game. Let ψ_i denote the payoff of bidder i in the Shapley value of game w. Surprisingly, perfect nesting under complete information yields the Shapley value for the game:

THEOREM 2: $\psi_i = x_i \; \forall i$

PROOF:

See Appendix, Part A.

TABLE 2—THE FAIR ALLOCATION RULE

Q	$x_1(Q)$	$x_2(Q)$	$x_3(Q)$
0	5	0	0
{1,2}	7	2	0
{1,3}	5	0	0
{2,3}	5	0	0
{1,2}{1,3}	8	3	1
{1,2}{2,3}	8	3	1
{1,3}{2,3}	7.33	2.33	2.33
{1,2,3}	8	3	1

Two interesting consequences of this theorem are worth noting. Characteristic function games with "conference structures" have been examined by Myerson (1980).[2] A conference is any set of two or more players who might meet together to discuss their cooperative plans. A conference structure, Q, then is any collection of conferences. Two players are connected in a given conference structure if they can be coordinated either by meeting together in some conference or by meeting in separate conferences that have one or more members in common to serve as intermediaries or by some longer sequence of overlapping conferences in the given conference structure. An allocation rule is a map that sends each possible conference structure, Q, into a feasible allocation of payoffs to players, $x_i(Q)$, $i: 1 \to n$. An allocation rule is fair if any two players always share equally the gain associated with their forming a conference. An allocation rule is stable if no player is ever harmed by the formation of a conference that includes him. Myerson's principal result is that there is a unique fair allocation rule and that this fair allocation rule for any given conference structure coincides with the Shapley value for the game, which is altered by requiring that players can only coordinate through the conferences in the given conference structure.[3] Myerson also shows that the fair allocation rule is stable if the game is superadditive.

The connection of Myerson's results to our own is immediate. Since the nested knockout imputation and the Shapley value of the complete information auction game coincide (Theorem 2) and since the characteristic function of this game is superadditive, we may conclude that the nested knockout imputation is consistent with the unique fair and stable allocation rule for the complete information auction game. This unique fair allocation rule is illustrated in Table 2 for the example given earlier.[4]

A fair allocation rule gives equal benefits to the members of each conference—all members of any given conference would gain equally from its formation or, equivalently, would lose equally from its dissolution. The gain, for instance, to forming {1,2} if no other conference exists is 4 in the example, and this is shared equally by players 1 and 2. Note that the allocation rule is stable—adding {1,2}, for example, to {1,3} and (2,3} helps both 1 and 2 (but hurts 3).

Overall, although there are other imputations in the core, the nested knockout impu-

[2] See also Myerson (1977). This paper examines "cooperation structures" in characteristic function games with transferable utilities. A cooperation structure is a special case of a conference structure in which every conference has exactly two members.

[3] Let S/Q denote the partition of S defined by the requirement that the subsets are the maximal connected coalitions that can be coordinated if the players meet only in the conferences of Q. Then the altered game is defined by

$$w(S/Q) \equiv \sum_{S_i \in S/Q} w(S_i) \qquad \forall S \subseteq N.$$

[4] The last conference structure listed, {1,2,3}, is equivalent to {{1,2},{1,3},{2,3}} in the sense that $S/\{\{1,2\},\{1,3\},\{2,3\}\} = S/\{1,2,3\}$ $\forall S \subseteq N$, where

$$N/Q \equiv \{\{j | i \text{ and } j \text{ are connected by } Q\} \, i \in N\}.$$

tation is the unique fair and stable allocation rule and, consequently, the Shapley value as well.

III. The Post-Auction Knockout

Now suppose that we relax the assumption that the valuations of members are public knowledge within the coalition and adopt instead the IPV framework in which v_p is private information to the pth bidder and modeled as an independent observation from the distribution function $F(v)$. It proves notationally convenient to assume that $F(v)$ can be characterized by a density function $f(v)$. We shall consider successively a simultaneous (single bid) and a sequential (multiple bid) version of the knockout. The equilibrium bids from these versions will be denoted, respectively, with superscripts "s" and "m." Throughout the discussion it is assumed that the ring initially acquires the item at price p_0, which is assumed to be less than v_1, the highest valuation among ring members.[5]

It will be shown below that ring members with valuations less than p_0 may nevertheless find it (*ex ante*) profitable to participate in the knockout as long as their valuations exceed a reservation value $v_* \leq p_0$ to be determined.

Following the seminal equivalence results in Myerson (1981), Milton Harris and Artur Raviv (1981) and, in particular, John Riley and William Samuelson (1981, Proposition 1), revenue comparisons between auction schemes essentially require the evaluation of the corresponding reservation values, which are defined as the valuations below which it is unprofitable to submit a bid. In particular, equality of the reservation values associated with two distinct auction schemes entails their revenue equivalence. Hence, the comparison below of the simultaneous and sequential knockouts are based entirely upon a comparison of the reservation values associated with each. Determining the relationship between these values is (nontrivially) established through a recursion argument. The bid functions that emerge from these derivations are of significant interest in their own rights.

A. *The Single Bid Knockout*

In this version of the knockout, members of the coalition are required to report a single bid to a "ring center." The ring members are then ordered by decreasing bids so that

$$(4) \qquad b_1 \geq b_2 \geq \cdots \geq b_n.$$

The bidder with the largest report wins the item, and the total payoff to the ring of $v_1 - p_0$ is allocated as follows. The winner receives

$$(5) \qquad x_1 = v_1 - b_2 + \sum_{p=2}^{n} \frac{b_p - b_{p+1}}{p},$$

and the rank l losing bidder, $l \neq 1$, receives

$$(6) \qquad x_l = \sum_{p=l}^{n} \frac{b_p - b_{p+1}}{p} \qquad l\colon 2 \rightarrow n,$$

where $b_{n+1} = p_0$ for notational convenience.

As in Section I, we shall assume that ring members who find it unprofitable to participate in the *ex post* knockout abstain from submitting bids. Hence, after these members have left the room, the n remaining *active* participants share the information that their n valuation are bounded below by $v_*(n, p_0)$, the reservation value associated with n and p_0, and for which we use the shorthand notation v_*. Let

$$F_p^l(v|v_*) = \sum_{j=0}^{p-1} \binom{l}{j} \times \frac{[1-F(v)]^j [F(v) - F(v_*)]^{l-j}}{[1-F(v_*)]^l}, \qquad p \leq l \leq n$$

[5] This assumption will be justified in Section IV where a mechanism will be identified that assures that ring will only acquire items at prices that do not exceed the highest valuation among its members. Here we focus upon the behavior of ring members within the ring—the behavior of the ring (and its members) within the larger auction is the subject of Section IV.

denote the distribution of the pth largest valuation among l valuations drawn independently from the distribution $F(\cdot)$, conditionally on them all being larger than the constant v_*. Let

$$f_p^l(v|v_*)=p\binom{l}{p}\times\frac{[1-F(v)]^{p-1}[F(v)-F(v_*)]^{l-p}}{[1-F(v_*)]^l}\times f(v)$$

denote the corresponding density. Under these circumstances

THEOREM 3: *The (symmetric) Nash equilibrium bidding strategy for a bidder with valuation v in a ring with n active members is to submit a bid $\zeta_n^s(v;v_*)$ satisfying*:

$$(7)\quad \zeta_n^s(v;v_*)=\frac{\int_v^\infty tf_2^n(t|v_*)\,dt}{1-F_2^n(v|v_*)},\qquad v\ge v_*,$$

where the reservation value v_* is zero if $p_0\le\zeta_n^s(0;0)$ and is otherwise given by the unique[6] *solution of the identity*

$$(8)\quad p_0=\zeta_n^s(v_*,v_*)=\int_{v_*}^\infty tf_2^n(t|v_*)\,dt$$

where p_0 is the price at which the ring acquired the item. Hence $v_*\le p_0$. The corresponding ex ante expected payoff to the bidder is

$$(9)\quad \tilde{\tilde{V}}_n^s(v|v_*)=\frac{b_*^s-p_0}{n}+\int_{v_*}^v(v-t)f_1^{n-1}(t|v_*)\,dt,$$

where $b_*^s=\zeta_n^s(v_*,v_*)$, so that $b_*^s=p_0$ if $v_*\ge 0$.

The proof of this proposition is available on request from the authors. An example is provided in the Appendix, Part B.

In words, the optimal bid for a player with valuation v is the expected value of the second highest valuation among n independent drawings from the distribution $F(\cdot)$ conditionally on (i) the n valuations being all larger than v_* and (ii) the second highest being larger than v. If the ring has acquired the item at a price $p_0\ge\zeta_n^s(0,0)$, then the *ex ante* expected payoff to a ring member with valuation v is given by the expectation of the function $\max\{0,v-t\}$, where t is the highest valuation among $n-1$ independent drawings from $F(\cdot)$, conditionally on them all being larger than v_*. If instead $p_0<\zeta_n^s(0,0)$, then participating ring members share an additional expected payoff given by $b_*{}^s-p_0$.

The Nash strategy (7) is monotone in v so that the ordering of the bids corresponds to that of the valuations and, in particular, the ring member with the largest valuation always wins the item. Furthermore, if $F_2^n(v|v_*)<1$, then

$$(10)\qquad \zeta_n^s(v;v_*)>v,$$

so that it is optimal for the ring members to overbid relative to their own valuations.[7] The

[6] The uniqueness of v_* follows from the fact that the partial derivative of $f_2^n(t|v_*)$ w.r.t v_* is negative for $t<t_*$ and positive for $t>t_*$, where $t_*\ge v_*$ is such that $nF(t_*)=(n-2)+2F(v_*)$. Since the integral of the derivative from v_* to ∞ is zero, it follows that the right-hand term in (8) is an increasing function of v_*. If, for example, valuations are uniformly distributed on [0,1], then

$$v_*=\begin{cases}0 & \text{if } p_0\le\dfrac{n-1}{n+1},\\[2ex] \dfrac{n+1}{2}\left[p_0-\dfrac{n-1}{n+1}\right]\le p_0 & \text{otherwise.}\end{cases}$$

[7] At a first price auction (IPV with homogeneous, risk neutral bidders), the symmetric Nash equilibrium bid strategy for a bidder with valuation v is equal to the

reason for doing so is obvious. A (monotone) strategy of underbidding cannot possibly be a Nash equilibrium since it implies that each term in the payoff functions (5) and (6) is nonnegative, including the term $v_1 - b_2$. Hence it would pay for a ring member to raise his bid at least up to his own valuation since by doing so he would increase his share of the ring total payoff even if he turned out to be the winner. Given this overbidding, ring members face the possibility that $v_1 - b_2$ might be negative. This restrains them from raising their bids arbitrarily since doing so would increase their chance of winning the item and, beyond a certain level, would only reduce their expected payoff.

Similar considerations will play an essential role in our analysis of the multiple bids knockout, where it will be shown that the sequential nature of the game makes it more attractive to risk higher bids during the earlier stages of the game—given the possibility of backing out at later stages—but at the same time raises the possibility that "loss-bidding" might occur at every stage of the game.

Note finally that decreasing n shifts the distribution function $F_2^n(t|v_*)$ to the left and, therefore, increases v_* as defined in (8). It follows that no ring member who has decided not to participate would change his mind on the basis of the corresponding decisions of the other bidders.

B. *The Multiple Bid Knockout*

We now consider a sequential game consisting of $n-1$ successive knockouts, each of which results in the elimination of one player from further contention. At each stage members submit bids with the difference between the smallest bid in the current stage and the smallest bid in the immediately preceding stage being divided equally among the members participating at the current stage. In addition, all except for the member submitting the smallest bid advance to the next stage. As usual within the context of sequential games, solutions are obtained by means of backward recursions. It proves notationally convenient to index the successive knockouts in reverse order and, more specifically, to use as index the number of players remaining in contention at the given stage of the game so that at stage l, the l remaining players are invited to submit bids and the one submitting the lowest bid is eliminated. During play of the game l runs backward from n to 1 and the last stage's winner gains possession of the item.[8] Given this indexing convention, the payoff functions to the players are still given by expressions (5) and (6) on the basis of the sequence of bids b_l^m $l\colon n \to 2$, where b_l^m denotes the lowest bid observed during round l. An important issue concerns the information made available to the remaining participants at the end of each round of the knockout. We adopt a stylized version of the actual bidding process in which inner ring bids are handled by a sole bidder and b_l^m is the only round l bid that is revealed to the participants.[9]

Let $\zeta_l^m(v)$ denote the round l (symmetric) Nash equilibrium strategy for the l players still in contention. Conceptually at least, ζ_l^m depends on the sequence of previously observed bids $b_{l+1}^m, \dots, b_n^m$. It will be shown, however, that ζ_l^m depends only upon v, is monotonically increasing, and thus can be

expected value of the second highest valuation from n given that the highest valuation equals v. The difference between this strategy and the one for the simultaneous knockout lies in the conditioning set. For the first price auction bidders attempt to win the item in order to obtain a positive surplus while at the simultaneous knockout it is advantageous to be the highest bidding loser.

[8] The bidders are also effectively indexed by l since, as we shall demonstrate below, by playing their sequential Nash equilibrium strategy, they are eliminated in increasing order of their valuations. Note, however, that the bids that are generated by the Nash strategies in the course of the game need no longer increase from one stage to the next. This is the relevance of "loss-bidding" in the present context.

[9] This assumption could be implemented by means of a thermometer device, such as that described in GM, which stops as soon as the lowest bidder withdraws and is reset at zero for each new round of the knockout.

inverted. Let $\lambda_l^m(b)$ denote the inverse function.

An important distinction emerges now that the bidding strategy changes at every round of the game. On one hand, the round l losing bid is given by

$$(11)\qquad b_l^m = \zeta_l^m(v_l) \qquad l: 2 \to n.$$

On the other hand, the $l-1$ players remaining in contention at the end of round l observe b_l^m and thus can base their next round decision on the information that their valuations $(v_1,\dots,v_{l-1})$ are bounded below by v_l, which they can retrieve from the inverse function $\lambda_l^m(b_l^m)$. Furthermore, since the Nash strategies are monotone increasing, it follows that the players' optimal bids at round $l-1$ are bounded below by b_l^*, where

$$(12)\qquad b_l^* = \zeta_{l-1}^m(v_l) \qquad l: 2 \to n.$$

Note that (12) also applies for $l = n+1$ with $v_{n+1} = v_*$ since at the opening of the knockout the n players have the information that the valuations are bounded below by the appropriate reservation value. Under these circumstances:

THEOREM 4: *The recursive symmetric Nash equilibrium bidding strategy for a bidder with valuation v who is active in round l is to submit a bid $\zeta_l^m(v)$ given by*

$$(13)\qquad \zeta_l^m(v) = \int_v^\infty t f_2^l(t|v)\,dt,$$

$$v \ge v_{l+1}, \qquad l: 2 \to n,$$

with a reservation value $v_{n+1} \equiv v_*$ as defined in (8). *The corresponding expected payoff for the rest of the game* (*including round l*) *is*

$$(14)\quad V_l^m(v|v_{l+1}) = \frac{b_{l+1}^* - b_{l+1}^m}{l} + \int_{v_{l+1}}^v (v-t) f_1^{l-1}(t|v_{l+1})\,dt,$$

where

$$b_{l+1}^m = \zeta_{l+1}^m(v_{l+1}), \qquad l: 2 \to n-1$$

$$b_{l+1}^* = \zeta_l^m(v_{l+1}), \qquad l: 2 \to n-1$$

$$b_{n+1}^m = p_0$$

$$b_{n+1}^* = b_*^s$$

with b_*^s as defined in Theorem 3.

The proof of this proposition is available on request from the authors. An example is provided in the Appendix, Part B. The critical fact that the two versions of the knockout share a common reservation value stems from the identity

$$(15)\qquad \zeta_n^m(v) \equiv \zeta_n^s(v; v).$$

In words, the optimal bid for a player with valuation v who is active at round l is the expected value of the second highest valuation among l independent drawings from the distribution $F(\cdot)$ conditionally on these l valuations all being larger than v. The expected payoff yet to accrue to the player is the expectation of the function $\max\{0, v-t\}$, where t is the largest valuation among $l-1$ independent drawings from $F(\cdot)$, conditionally on them all being larger than v_{l+1}, corrected by an equal share of the difference between the round l lower bound b_{l+1}^* and the previous round lowest effective bid b_{l+1}^m.

The interpretation of ζ_l^m in expression (13) indicates that

$$(16)\qquad \zeta_l^m(v) \ge \zeta_{l-1}^m(v)$$

for all v and, in particular, for $v = v_l$ so that

$$(17)\qquad b_l^m \ge b_l^* \qquad l: 2 \to n.$$

Furthermore, since ζ_{l-1}^m is monotone increasing, we also have

$$(18)\qquad b_{-1}^m \ge b_l^* \qquad l: 2 \to n.$$

However, as the example provided in the Appendix, Part B, illustrates, it is no longer the case that b_{l-1}^m is necessarily larger than b_l^m, hence the possibility of "loss-bidding."

The comparison between the sequential game and the single bid game is now straightforward.

COROLLARY 5: *Under the conditions of Theorems* 3 *and* 4

$$(19)\quad \tilde{V}_n^s(v|v_*) = \tilde{V}_n^m(v|v_*), \qquad \forall v \geq v_*$$

$$(20)\quad F_2^n(v|v_*) > 0 \Rightarrow \zeta_n^m(v) > \zeta_n^s(v)$$

$$(21)\quad \tilde{V}_l^s(v|v_{l+1}) \geq \tilde{V}_l^m(v|v_{l+1}) \qquad l: 2 \rightarrow n-1$$

PROOF:

(19) follows from Riley-Samuelson (1981, Proposition 1) given that the corresponding reservation values are equal; compare (7) with (14) for $l = n$. (20) follows from our interpretation of the results in expressions (7) and (13). (21) follows from the inequality in expression (17). □

The revenue equivalence (19) implies, but is stronger than, the obvious result that unconditionally on v, the *ex ante* payoffs are the same in both versions—a necessary consequence of the homogeneity of the players. Trivially, the latter imputation also corresponds to the Shapley value for the game with characteristic function:

$$(22)\quad w(S) = E\left[\max\left\{\max_{p\in S} v_p - \max_{q\notin S} v_q, 0\right\}\right] \qquad \forall S \subseteq N,$$

where E denotes the expectation operator with respect to all valuations.

Expression (20) demonstrates that the sequential version of the game generates additional overbidding compared to the single bid version in the opening round. Essentially, the possibility of revising one's bid during the later rounds makes it more attractive to gamble for a larger first-round payoff. Expression (21) indicates, however, that after the first round the multiple bid game becomes less attractive to the remaining players since they are now confronted at every round with the possibility of loss bidding. It is this second effect that prevents a player from arbitrarily raising his opening bid in the sequential game. Our equivalence result (19) indicates that the two effects exactly offset each other *ex ante* relative to the single bid version of the game. Graphs of the bid functions for specific examples are found in the Appendix, Part B.

IV. The Private Information Auction Game

To this point we have demonstrated that the nested knockout has unique properties in a complete information environment (Section II), but that with incomplete information there is systematic overbidding (Section III). The latter result was derived under the assumption of *ex ante* bidder homogeneity—ring members draw their valuations from the *same* distribution. Given that nesting is used in practice when ring members differ from one another *ex ante*—draw valuations from different distributions—this leaves an important question unanswered. What results hold for heterogeneous bidders under incomplete information? In this section we propose a mechanism that a coalition of *ex ante* heterogeneous bidders at a second price auction can employ that induces ring members to report truthfully their valuations to the ring, assures that the highest reporting ring member will win an item obtained by the ring, pays members the Shapley imputation, and creates no incentives for cheating.

Suppose now that we extend the IPV model to allow for heterogeneous bidders by modeling the valuation of the ith bidder to be an observation from $F_i(v)$ for $i = 1,2,\ldots,n$. These distribution functions are assumed to be common knowledge to all bidders. For expositional convenience we retain the assumption that the auctioneer sets a reserve price of zero.

As noted in Section III the post-auction knockout does not provide incentives for ring members to report their valuations truthfully. This represents no difficulty for the case of homogeneous bidders since the monotonically increasing and symmetric bid strategy identified there assures that the bidder with the highest valuation will ultimately win the item. Under heterogeneity, however, such a knockout gives rise to asymmetric bid strategies and positive probabilities of the item not being won by the bidder with the highest valuation. Such potential inefficiencies reduce the expected payoff to ring formation and motivate the search for more efficient mechanisms.

Consider a mechanism that requires the individual members of the ring to make prior, not necessarily truthful, reports regarding their private valuations to the ring-center. Then for any given vector of reports of private valuations from ring members, the mechanism must determine, perhaps randomly: (1) the recommended bid for each member to submit at the main auction, (2) the ring member that will ultimately receive an item that is won by the ring at the main auction, (3) the payments to collect from (make to) each member of the ring.

Such a mechanism is called incentive compatible if and only if it is a Nash equilibrium in the resulting game for each member to participate (the expected payoff to each member in this equilibrium is at least as great as the payoff that member could expect by not participating), to report his private valuation truthfully, and to submit the recommended bid at the main auction. A given mechanism dominates another if the expected payoff to each and every ring member is at least as great in the former mechanism as in the latter one, regardless of the ring members' private valuations. An incentive compatible mechanism is called durable if the members of the ring would never unanimously approve a change to another mechanism even if they knew more than just their own valuations, that is, communication had occurred—see Holmström and Myerson (1983).

With these preliminaries in mind consider the following mechanism, which we call the second price pre-auction knockout or PAKT:

> The "ring center" makes a fixed payment, P_i, to each of the ring members prior to the main auction. (The magnitude of these payments will be identified shortly). Each of the n members of the ring submits a sealed "reported bid" to the ring center at this time. The member submitting the highest bid is selected by the ring center as the sole bidder and advised to bid on his own behalf at the main auction. Ring members other than the sole bidder are advised not to submit bids at the main auction. Should the sole bidder win the item at the main auction, he would pay the auctioneer the contracted price and, additionally, would pay the ring center the difference between the second highest reported bid from the ring and the contracted price provided that this difference is positive.

This scheme assures the winning bidder of receiving a guaranteed fixed payment and paying a price equal to the second highest of all bids and reported bids for the item. Part of this price goes to the auctioneer and part to the ring center—one requirement for identifying the fixed payments will be that their sum equals the expected value of that portion of the price paid to the ring center. The ring center serves in this context as both "mediator" and "budget-breaker"; see Holmström (1982). Our interest in the second price PAKT stems from the following proposition.

THEOREM 6: *Truthful reporting is a dominant strategy for participants in the second price PAKT.*

PROOF:

See Appendix, Part C.

One obvious implication of this result is that the bidder with the highest valuation from the ring will report this valuation to the ring center and subsequently submit this valuation as a bid. Thus no bidder, ring or non-ring, will ever pay more for the item than his or her valuation.

Given the dominant strategies of truthful reporting to the ring-center by members of S and to the auctioneer by members of $N-S$, we can again reasonably employ the definition of the characteristic function given by equation (3) with the understanding that the expectations operator now accounts for the heterogeneous distributions across bidders. Let γ_i equal the expected value of the difference between the highest valuation and the second highest valuation from the ring given that the highest valuation belongs to the ith bidder. As before, let ψ_i be the Shapley imputation to the ith bidder. Then define the fixed payments for the second price PAKT as follows:

$$P_i \equiv \psi_i - \gamma_i. \tag{23}$$

With these fixed payments the second price PAKT yields an imputation identical to the Shapley value. We also note

THEOREM 7: *With fixed payments corresponding to* (23), *the second price PAKT is an incentive efficient and durable mechanism.*

PROOF:
See Appendix, Part C.

Note also that the results of Myerson discussed in Section III again imply that the allocation resulting from the second price PAKT is consistent with the requirements for both fair and stable allocation rules.[10]

V. Conclusion and Future Research

Within the independent private values model the total amount available to a bidder coalition to divide among its members at either a second price or English auction is equal to the difference between the maximal valuation within the ring and the price paid for the item at the main auction, given that the difference is positive. The central issue of this paper has been to determine the characteristics of the mechanism used by practicing bidder coalitions to disperse this difference among its members.

In order to obtain an initial understanding of the nested knockout, several special cases have been studied. The decomposition of the problem involves three main issues. First, payments to ring members could be made *ex ante* (before valuations are drawn by members) or *ex post*. Second, the bidders could be homogeneous or heterogeneous in terms of their distributional identity. Third, with respect to valuations drawn within the ring, the auction could be modeled as a game of complete or incomplete information.

Graham and Marshall (1987) have already studied the *ex ante*, homogeneous, incomplete information environment. In this study we find that the payments to ring members described by Graham and Marshall (1987) are the *ex ante* Shapley values of the bidders.

In an *ex post*, heterogeneous, complete information auction game, it has been demonstrated that the nested knockout results in side payments to ring members that are not only in the core but exactly equal to their Shapley values. Within a more realistic setting, the incomplete information auction game, it has been shown that homogeneous ring members overbid in both the simultaneous and sequential versions of the *ex post* nested knockout. These results demonstrate the incentive compatibility problems of the nested knockout. Finally, in the incomplete information auction game with heterogeneous bidders, an *ex ante* nested knockout allocation mechanism that awards ring members their *ex ante* Shapley values has been proposed. This mechanism also has the desirable characteristic of being both incentive efficient and durable (in the sense of Holmström and Myerson, 1983).

Although we have posed an *ex ante* mechanism in Section IV that achieves efficiency subject to incentive compatibility constraints, the most pressing issue for future research concerns the *ex post* use of a nested knockout by heterogeneous ring members in the incomplete information auction game. The main problem is to determine whether potential inefficiencies within the ring are unavoidable. Specifically, does there exist Nash equilibrium behavior that precludes the

[10] The susceptibility of both second price and English auctions to stable collusive behavior has been studied in GM.

possibility of a ring member without maximal valuation from winning the object? If this possibility is not eliminated, then not only will the Shapley value be unattainable as a characterization of the resulting side payments to ring members but the coalition will not be maximizing its total payoff. This case is of particular interest since it most closely parallels the reality of practicing bidder coalitions.

Another issue for future research involves the characterization of the nested knockout within a common values or general symmetric model setting. Bidder coalitions may not make use of *ex ante* payment schemes because of the value of the information that is transmitted to members during the bidding at the main auction. Specifically, *ex ante* bidding commitments (i.e., no one bids but the highest valued ring member) may not be self-enforcing within such a setting.

APPENDIX

A. *Results for the Complete Information Auction Game*

PROOF OF THEOREM 1:[11]

Recall that the valuations are ordered:

$$v_1 \geq v_2 \geq \cdots \geq v_n$$

and the nested knockout imputation for the ring is

$$x_i = x_{i+1} + \frac{v_i - v_{i+1}}{r_i},$$

where r_i denotes the number of bidders with a valuation at least as great as the ith bidder. We need to show

$$\sum_{i \in S} x_i \geq w(S) \qquad \forall S \subseteq N.$$

Since $x_i \geq 0\ \forall i$ and

$$w(S) = \begin{cases} 0 & \text{if } 1 \notin S \\ v_1 - v_k & \text{if } 1,2,\ldots,k-1 \in S, \\ & \quad \text{and } k \notin S \end{cases}$$

we are done if $1 \notin S$. Suppose then that $1,2,\ldots,k-1 \in S$ and $k \notin S$. Then

$$\sum_{i \in S} x_i = \sum_{i=1}^{k-1} x_i + \sum_{\substack{i \in S \\ i > k}} x_i$$

$$= v_1 - v_k + \sum_{\substack{i \in S \\ i > k}} x_i$$

$$\geq v_1 - v_k.$$

PROOF OF THEOREM 2:

Let ψ_i^p denote the Shapley imputation to player i in the all inclusive, complete information auction game among players $\{1,2,\ldots,p\}$ with valuations $v_1 \geq v_2 \geq \cdots \geq v_p$, where $p = 1,2,\ldots,n$. We consider the addition of bidders to this all inclusive game. Consider now the recursion from $p-1$ to p. Bidder p comes in with the lowest valuation among the p players. He can enter in p different positions in each permutation of the $p-1$ other bidders. Only if he is last is his marginal contribution to the coalition positive. This occurs $(p-1)!$ times, of which $(p-2)!$ are against bidder i, $i = 1,2,\ldots,p-1$, in which case player i's marginal contribution is reduced by v_p. Hence

$$\psi_i^p = \psi_i^{p-1} - \frac{(p-2)!}{p!} v_p \qquad i = 1,2,\ldots,p-1$$

$$\psi_p^p = \frac{(p-1)!}{p!} v_p$$

$$= \frac{v_p}{p}.$$

It follows from recursion that

$$\psi_i^p = \frac{v_i}{i} - \frac{v_{i+1}}{(i+1)i} - \frac{v_{i+2}}{(i+2)(i+1)} - \cdots - \frac{v_p}{p(p-1)}$$

$$= \frac{v_i - v_{i+1}}{i} + \frac{v_{i+1} - v_{i+2}}{i+1} + \cdots + \frac{v_p}{p}$$

and thus that

$$\psi_i^p = \sum_{j=i}^{p-1} \frac{v_j - v_{j+1}}{j} + \frac{v_p}{p} \qquad i = 1,2,\ldots,p-1.$$

This is precisely the knockout imputation when evaluated at $p = n$.

B. *The Post Auction Knockout: An Example*

The proofs of the propositions of Section III are somewhat lengthy and have been omitted. They are, of

[11] It could, alternatively, be shown that the game w is convex. Theorem 1 would then follow from this fact and Theorem 2.

TABLE 3—CONDITIONAL DENSITY FUNCTION

	$\Pr(E_i)$	$\Pr(E_i \vert v)$	$f(v \vert E_i)$	$f(v_a \vert v, E_i)$	$f(v_b \vert v, E_i)$
E_1	1/3	v^2	$3v^2 I_{(0,1)}$	$\frac{2v_a}{v^2} I_{(0,v)}$	$\frac{2(v-v_b)}{v^2} I_{(0,v)}$
E_2	1/3	$2v(1-v)$	$6v(1-v)I_{(0,1)}$	$\frac{1}{1-v} I_{(v,1)}$	$\frac{1}{v} I_{(0,v)}$
E_3	1/3	$(1-v)^2$	$3(1-v)^2 I_{(0,1)}$	$\frac{2(v_a-v)}{(1-v)^2} I_{(v,1)}$	$\frac{2(1-v)}{(1-v)^2} I_{(v,1)}$

course, available upon request from the authors. These results can be illustrated, however, by means of a simple example where our focal player with valuation v faces two other players with ordered valuations $V_a > V_b$. The three valuations are drawn independently from the uniform distribution on [0,1]. It is further assumed that $p_0 = v_4 = 0$.

Since the expectation of the highest of the three valuations is 3/4, it follows that the *ex ante* Shapley value for each of the three players is 1/4.

Three events are considered in the discussion that follows. Let E_i denote the event that the focal player is ith in the ranking, where i: 1 → 3. Conditionally on E_i, the density of V and those of V_a and V_b given that $V = v$ are given in Table 3, where $I_{(a,b)}$ denotes the indicator function for the interval $]a,b[$.

We will now discuss in turn the (sequential) multiple-bid and the single-bid versions of the three-player knockout.

The Multiple Bid Game. The first ($l = 3$) and second ($l = 2$) Nash equilibrium bid functions (13) are respectively given by

$$\zeta_3^m(v) = \tfrac{1}{2}(1+v), \tag{24}$$

$$\zeta_2^m(v) = \tfrac{1}{3}(1+2v).$$

Their graphs are given by the two lines in Figure 1. For comparison, the curve illustrates the Nash equilibrium strategy for the single bid game.

Let $x = \psi(v, V_a, V_b)$ denote the (random) payoff to our focal player with valuation v. Conditionally on E_i, i: 1 → 3, x is given by

$$E_1\colon\ x = v - \tfrac{1}{2}\zeta_2^m(V_a) - \tfrac{1}{6}\zeta_3^m(V_b), \tag{25}$$

$$E_2\colon\ x = \tfrac{1}{2}\zeta_2^m(v) - \tfrac{1}{6}\zeta_3^m(V_b),$$

$$E_3\colon\ x = \tfrac{1}{3}\zeta_3^m(v).$$

On the basis of the density functions of V_a and V_b, as given in Table 3, we can evaluate the expectation of x, conditional on E_i and v or conditional on E_i only, and find that

$$E(x|E_1, v) = \tfrac{1}{4}(3v-1) \tag{26}$$

$$E(x|E_1) = \tfrac{5}{16} \approx 0.313$$

$$E(x|E_2, v) = \tfrac{1}{24}(7v+2)$$

$$E(x|E_2) = \tfrac{11}{48} \approx 0.229$$

$$E(x|E_3, v) = \tfrac{1}{6}(v+1)$$

$$E(x|E_3) = \tfrac{5}{24} \approx 0.208.$$

Unconditionally on E_i, we find that

$$E(x|v) = \tilde{V}_3^m(v|0) \tag{27}$$

$$= \tfrac{1}{6}(1+2v^3).$$

The expectation of the latter expression over V obviously coincides with the *ex ante* Shapley value of 1/4.

The graphs of the expectations of x conditionally and unconditionally on E_i are given in Figure 2. The unconditional expectations of the 1st, 2nd, and 3rd highest valuations are respectively 0.75, 0.50, and 0.25. Note that the ordering between the conditional expectations $E(x|E_i, v)$ critically depends on v and reflects the importance of loss-bidding in the second round. A player, for example, with valuation $v < 2/3$ would benefit from being eliminated in the first round since he can only lose in expected payoff by surviving the round.[12]

The Single Bid Game. The single bid Nash equilibrium strategy (7) is given by

$$\zeta_3^s(v;0) = \frac{1}{2}\,\frac{1+2v+3v^2}{1+2v}.$$

[12] At $v = 2/3$, $E(x|E_2, v) = E(x|E_3, v)$.

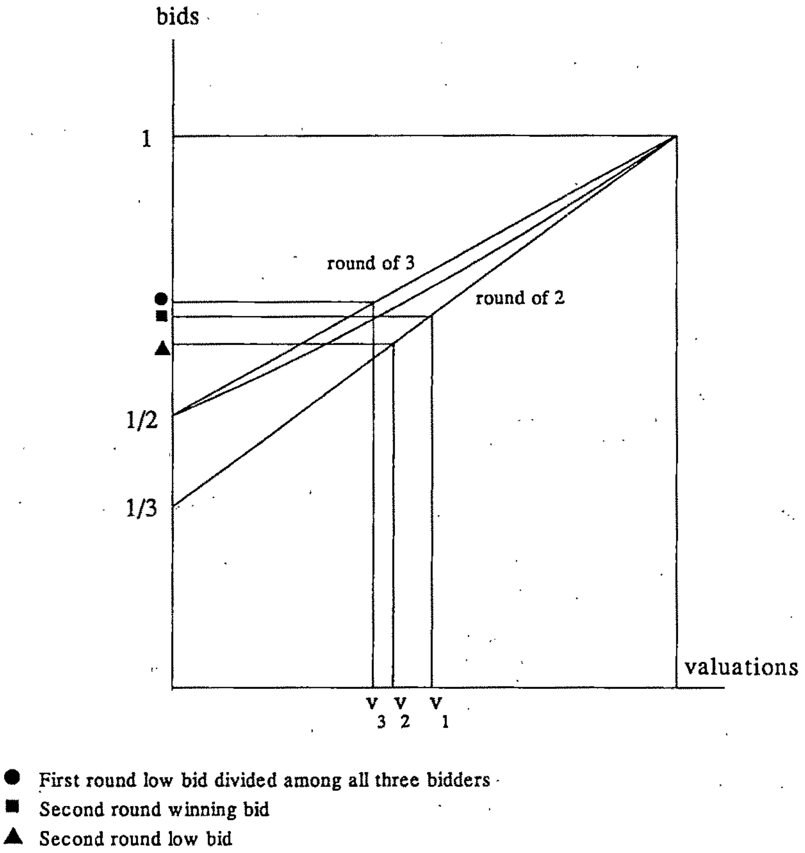


FIGURE 1. EQUILIBRIUM STRATEGIES IN THE SINGLE- AND MULTIPLE-BID GAMES

The graph of $\zeta_3^s(\cdot)$ is illustrated in Figure 1. Note that

$$\zeta_3^m(v;0) > \zeta_3^s(v;0) > \zeta_2^m(v;0) \qquad \forall v \in (0,1), \tag{28}$$

which nicely illustrates our earlier discussion of the two games.

The conditional payoffs are still given by equation (25) except that the bid functions ζ_2^m and ζ_3^m are now replaced by ζ_3^s. A conceptually straightforward though somewhat tedious derivation leads to the following expressions for the conditional payoffs.

$$E(x|E_1,v) = \frac{1}{48}\left[34v - 4 - \frac{6}{v} + 3\frac{1-v}{v^2}\ln(1+2v)\right],$$

$$E(x|E_2,v) = \frac{1}{4}\frac{1+2v+3v^2}{1+2v} - \frac{1}{96}\left[6v + 2 + \frac{3}{v}\ln(1+2v)\right],$$

$$E(x|E_3,v) = \frac{1}{6}\frac{1+2v+3v^2}{1+2v},$$

$$E(x|E_1) = \frac{7}{96} + \frac{27}{128}\ln 3 \approx 0.305,$$

$$E(x|E_2) = \frac{91}{96} - \frac{81}{128}\ln 3 \approx 0.253,$$

$$E(x|E_3) = -\frac{13}{48} + \frac{27}{64}\ln 3 \approx 0.193.$$

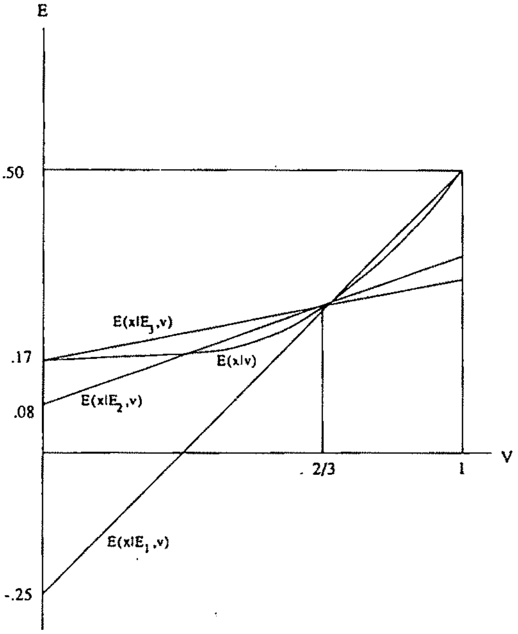


FIGURE 2. EXPECTED PAYOFF IN THE MULTIPLE-BID GAME

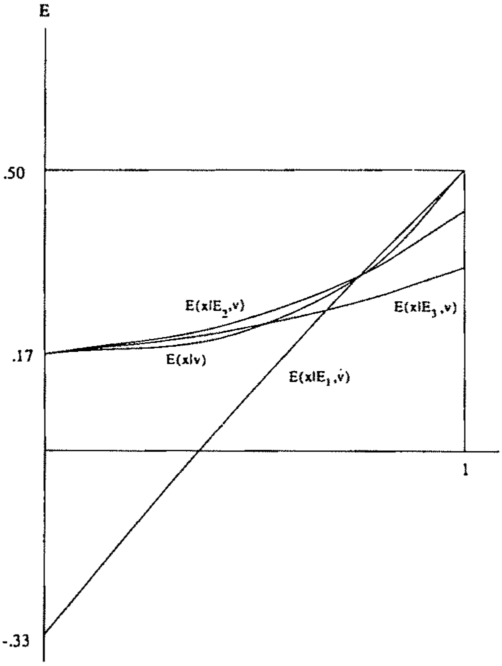


FIGURE 3. EXPECTED PAYOFFS IN THE SINGLE-BID GAME

Note also that $E(x|v)$ coincides with the expression in (27) as established in Corollary 5. The graphs of the expectations of x, conditionally and unconditionally on E_i, are found in Figure 3.

C. *Results for the Private Information Auction Game*

PROOF OF THEOREM 6:

In the event that the ring wins the item, the member submitting the highest reported bid is awarded the item and pays a total price equal to the second highest bid/reported bid of the $N+1$ bids ($N-K$ non-ring bids at the main auction, K reported bids from the ring members and the reserve price bid of the auctioneer). The payment to each ring member is a fixed non-contingent constant and therefore cannot affect incentives. Since the second price PAKT requires the winning bidder to pay the second price of all bidders and payments to ring members do not affect incentives, it follows from the logic of Vickrey's (1961) seminal paper that it is a Nash equilibrium strategy for each member of the ring to report his or her valuation truthfully to the ring-center and to follow his bidding recommendations at the main auction. In fact, truthful revelation and compliance is not only a Nash strategy, it is a dominant strategy as well.

PROOF OF THEOREM 7:

Voluntary participation is also advantageous. Ring membership entails three possible outcomes for a given ring member. First, if the ring does not acquire the item, membership is advantageous since the ring member receives P_i and would have obtained nothing acting individually. Second, if the ring wins the item but the item is awarded to another member, membership is still advantageous since once again the member receives P_i and would have obtained nothing acting individually. Third, if the ring wins the item and it is awarded to the ring member then membership is still advantageous since the number pays precisely the same price as would have been necessary acting individually but again receives P_i. Consequently, voluntary participation is also a dominant strategy.

The equilibrium strategies assure that the ring will win the item in every circumstance in which some member has a valuation exceeding the cost of acquiring the item. Further, in every case in which the ring wins the item it is awarded to that member with highest valuation. The mechanism, in short, assures the ring members of the greatest possible joint payoff in each contingency. Consequently, there exists no other "balanced budget" mechanism, whether incentive compatible or not, that dominates the second price PAKT. Therefore, the second price PAKT is incentive efficient.

Since the second price PAKT is incentive efficient and since participation and truthful revelation of valuations in the second price PAKT is a dominant strategy it follows that the second price PAKT is a durable mechanism (by Theorem 2 of Holmström and Myerson, 1983).

REFERENCES

Cassady, Ralph, Jr., *Auctions and Auctioneering*, Berkley: University of California Press, 1967.

Graham, Daniel A. and Marshall, Robert C., "Collusive Behavior at a Single Object English Auction," Duke University, Department of Economics Working Paper No. 85-01, 1985.

______ **and** ______, "Collusive Behavior at Single Object Second Price and English Auctions," *Journal of Political Economy*, December 1987, *95*, 1217–39.

Harris, Milton and Raviv, Artur, "Allocation Mechanisms and the Design of Auctions," *Econometrica*, November 1981, *49*, 1477–99.

Holmström, Bengt, "Moral Hazard in Teams," *Bell Journal of Economics*, Autumn 1982, *12*, 324–40.

______ **and Myerson, Roger B,** "Efficient and Durable Decision Rules with Incomplete Information," *Econometrica*, November 1983, *51*, 1799–819.

Myerson, Roger B, "Graphs and Cooperation in Games," *Mathematics of Operations Research*, August 1977, *2*, 225–29.

______, "Conference Structures and Fair Allocation Rules," *International Journal of Game Theory*, March 1980, *9*, 169–82.

______, "Optimal Auction Design," *Mathematics of Operations Research*, February 1981, *6*, 58–73.

______, "Basic Theory of Optimal Auctions," in Martin Schubik and Richard Englebrecht-Wiggans, eds., *Auctions, Bidding and Contracting: Uses and Theory*, New York: New York University Press, 1983.

Riley, John and Samuelson, William, "Optimal Auctions," *American Economic Review*. June 1981, *71*, 381–92.

Vickrey, William, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance*, March 1961, *16*, 8–37.

The Economics of Modern Manufacturing: Technology, Strategy, and Organization

By PAUL MILGROM AND JOHN ROBERTS*

Manufacturing is undergoing a revolution. The mass production model is being replaced by a vision of a flexible multiproduct firm that emphasizes quality and speedy response to market conditions while utilizing technologically advanced equipment and new forms of organization. Our optimizing model of the firm generates many of the observed patterns that mark modern manufacturing. Central to our results is a method of handling optimization and comparative statics problems that requires neither differentiability nor convexity. (JEL 022)

In the early twentieth century, Henry Ford revolutionized manufacturing with the introduction of his "transfer line" technology for mass production, in which basic inputs are processed in a fixed sequence of steps using equipment specifically designed to produce a single standardized product in extremely large quantities for extended periods of time. Although the specialization of Ford's factories was extreme—the plants had to be shut down and redesigned when production of the Model T was ended—the transfer line approach influenced generations of industrialists and changed the face of manufacturing (see David A. Hounshell, 1984).

In the late twentieth century, the face of manufacturing is changing again.[1] First, the specialized, single-purpose equipment for mass production which had characterized Ford's factories is being replaced by flexible machine tools and programmable, multitask production equipment. Because these new machines can be quickly and cheaply switched from one task to another, their use permits the firm to produce a variety of outputs efficiently in very small batches,[2] especially in comparison to the usual image of mass production (Nicholas Valery, 1987). Kenneth Wright and David Bourne (1988) report that in a recent survey of aerospace and other high precision industries 8.2 percent of all batches were of size one and 38 percent were sixteen or less. An Allen-Bradley Company plant making electric controls is reported to be able to switch production among its 725 products and variations with an average changeover time for resetting equipment of six seconds, enabling it to schedule batches of size one with relative efficiency (Tracy O'Rourke, 1988). Even in the automobile industry, flexible equipment has become much more common. Recently,

*Professor of Economics, Department of Economics. Stanford University, Stanford, CA. 94305-6072 and Jonathan B. Lovelace, Professor of Economics, Graduate School of Business, Stanford University, CA. 94305-5015. The research reported here was supported by the National Science Foundation and the Center for Economic Policy Research of Stanford University. We are grateful to Tim Bresnahan, Morris Cohen, Xavier de Groote, Victor Fuchs, John McMillan, Roger Noll, Mike Riordan, Nate Rosenberg, Ed Steinmueller, Hal Varian, Steve Wheelwright, and two anonymous referees for helpful discussions, comments, and suggestions.

[1] Probably no single firm is involved in all the changes we will describe. Nevertheless, there is a definite, discernable pattern of change in technology, manufacturing, marketing, and organizational strategy that characterizes successful "modern manufacturing." For a description of the technologies involved, see U.S. Congress Office of Technology Assessment, 1984.

[2] Optimal batch size can be determined via a standard Economic Order Quantity model, in which the setup costs of switching from making one product to making another are traded off against the costs of holding the larger average inventories of finished goods that go with longer runs and less frequent changeovers. Optimal batch size is a decreasing function of setup costs and so batch sizes optimally decrease as more flexible machines are introduced.

General Motors' engineers were able, for the first time in company history, to use a regular, producing facility to make pilots of the next year's model cars. The engineers set the equipment to make 1989 models after workers left the factory on Friday afternoon, ran the equipment to manufacture the new models over the weekend, and then reset the equipment to produce 1988 model cars so that regular production could be resumed on Monday morning (Thomas Moore, 1988). In contrast, with the older, less flexible technologies that have been the norm in the industry, changing over to produce the new year's models typically involved shutting down production for weeks.

Flexible equipment and small batch sizes have been accompanied by other changes. Smaller batch sizes are directly associated with a shortening of production cycles and with reductions in work-in-process and finished goods inventories. Shorter product cycles in turn support speedier responses to demand fluctuations and lead to lower back orders. Indeed, a general strategic emphasis on speeding up all aspects of the firm's operations is becoming common (Brian Dumaine, 1989). This is manifested in shorter product-development times, quicker order-processing, and speedier delivery, as well as in producing products faster. Examples abound. General Electric has reduced the design and production time it takes to fill an order for a circuit-breaker box from three weeks to three days, in the process reducing back orders from sixty days to two (Dumaine). The Allen-Bradley plant mentioned above fills orders the day after they are received, then ships them that same day by air express (O'Rourke). Building on early development by Toyota, many manufacturers now plan production jointly with their suppliers and maintain constant communication with them. This allows the downstream firms to replace inventories of components and supplies with "just-in-time" deliveries of needed inputs (James Abegglen and George Stalk, Jr., 1985; Nicholas Valery, 1987). Combining flexible production and low finished goods inventories with reliance on electronic data communications, Benetton maintains inventories of undyed clothing (shirts, scarves, pullovers) and uses nightly sales data gathered at its automated distribution center from terminals in individual stores to determine the colors it should make and where to ship its output (Tom Peters, 1987). Strategies like this one require not only flexible equipment to produce the right products, but short production cycles so that products are available at the right time. The extreme of this line of development is production of previously mass-produced items on a make-to-order basis: Moore reports a widespread rumor that the GM Saturn project will involve cars being custom-built within days of receipt of customers' computer-transmitted individual orders.

The manufacturing firms that adopt these new technologies and methods appear to differ from traditional firms in their product strategies as well. Many firms are broadening product lines, and there is a widespread increased emphasis on quality, both through frequent product improvements and new product introductions, and through reductions in defects in manufacturing. Caterpillar Corporation's $1.2 billion "Plant with a Future" modernization program has been accompanied by a doubling of the size of its product line (Ronald Henkoff, 1988). Rubbermaid insists that 30 percent of its sales should come from the products introduced in the preceding five years, and 3M Company has a similar 25 percent rule for each of its 42 divisions, with 32 percent of its $10.6 billion in 1988 sales actually coming from products less than five years old (Russell Mitchell, 1989). Meanwhile, reports of order-of-magnitude reductions in percentage defects are becoming commonplace.

New organizational strategies and workforce management policies are also part of this complex of changes. Ford has adopted a parallel, team (rather than sequential) approach to design and manufacturing engineering that, in conjunction with CAD/CAM (*Computer Aided Design/Computer Aided Manufacturing*) techniques, has cut development time on new models by one third (Alex Taylor, III, 1988). AT&T successfully used a similar, multidepartmental team approach in developing its new 4200 cordless phone (Dumaine), as did NCR with its recently introduced 2760 electronic cash register (Otis Port, 1989b). Lockheed Corp.'s Aeronautical Systems Group has managed to reduce the time for designing and manufacturing sheet-metal parts by 96 percent, from 52 days to 2; the project manager credits organizational changes (including the arrangement of workstations, redefinition of worker responsibilities, and adoption of team approaches) with 80 percent of the productivity gain (Port, 1989a, p. 143; Warren Hausman, 1988). Motorola's adoption of a pay scheme based on the skills employees acquire (rather than on their job assignments), its elimination of segmented pay categories among production workers, and its giving workers multiple responsibilities (including having production workers do quality inspections) is credited with major improvements in quality (Norm Alster, 1989). A further auto industry example comes from GM's massive investments in new technology, which have gone hand-in-hand with new supplier relations and more flexible work arrangements, as well as a broadened product line (General Motors Corporation, 1988).

More generally, Michael J. Piore (1986) provides survey evidence from firms around Route 128 in Boston of wider product lines, shorter product life cycles, greater emphasis on product quality, increased reliance on independent suppliers and subcontractors, and a more flexible organization of work that is supported by new compensation policies. Banri Asanuma, (1988a), has found similar trends among Japanese firms, Valery provides more anecdotal evidence drawn from a wide variety of industries internationally, and earlier Piore and Charles F. Sabel (1984) described related developments among small businesses in Italy and Austria.

A striking feature of the discussions of flexible manufacturing found in the business press is the frequency with which it is asserted that successful moves toward "the factory of the future" are not a matter of small adjustments made independently at each of several margins, but rather have involved substantial and closely coordinated changes in a whole range of the firm's activities. Even though these changes are implemented over time, perhaps beginning with "islands of automation," the full benefits are achieved only by an ultimately radical restructuring. Henkoff (p. 74) noted that one of the lessons of Caterpillar's program was: "Don't just change selected parts of your factory, as many manufacturers have done. To truly boost efficiency...it's necessary to change the layout of the entire plant." The first lesson that Dumaine drew from studying successful adoption of speed-based strategies was to "start from scratch." In discussing the adoption of "computer integrated manufacturing" (CIM), Valery (p. 15) stated that "nothing short of a total overhaul of the company's strategy has first to be undertaken." And in a parallel fashion, Walter Kiechel III (1988, p. 42) noted: "To get these benefits (of more timely operations), you probably have to totally redesign the way you do business, changing everything from procurement to quality control."

This paper seeks to provide a coherent framework within which to understand the changes that are occurring in modern manufacturing. We ask, Why are these changes taking place? Is it mere coincidence that these various changes appear to be grouped together, or is there instead some necessary interconnection between them and common driving force behind them? What are the implications of the changes in manufacturing technology for inventory policy, product market strategy, and supplier and customer relations? What are the implications for the "make or buy" and vertical integration decisions and for the structure of business organization more generally?

Our approach to these questions is a price-theoretic, supply-side one involving three elements: exogenous input price changes, complementarities among the elements of the firm's strategy, and non-convexities. The first element is the effect of technological change in reducing a set of costs. The particular ones on which we focus include: the costs of collecting, organizing, and communicating data, which have been reduced over time by the development of computer networks and electronic data transmission systems; the cost of product design and development, which have fallen with the emergence of computer-aided design; and the costs of flexible manufacturing, which have declined with the introduction of robots and other programmable production equipment. We take these relative price reductions, whose existence is well documented, to be exogenous.

The direct effect of any of these price changes individually would be to increase use of the corresponding factor: for example, the emergence (i.e., falling cost) of CAD/CAM encourages its adoption, and the reduced cost of designing and beginning production of new products directly increases the attractiveness of expanded product lines and frequent product improvements. However, with several relative prices falling, there are multiple interactions, both among the corresponding technological factors and between them and marketing and organizational variables. These interactions give rise to indirect effects that might in principle be as large as the direct effects, and opposite in sign. Here, the second element of our analysis appears: These indirect effects tend, in the main, to reinforce the direct effects because the corresponding relationships are ones of "complementarity." Here, we use the term "complements" not only in its traditional sense of a relation between *pairs of inputs*, but also in a broader sense as a relation among *groups of activities*. The defining characteristic of these groups of complements is that if the levels of any subset of the activities are increased, then the marginal return to increases in any or all of the remaining activities rises. It then follows that if the marginal costs associated with some activities fall, it will be optimal to increase the level of all of the activities in the grouping.

As an illustration, let us trace some of the indirect effects of a fall in the cost of computer-aided design (CAD) equipment and software that leads to the equipment being purchased. Some CAD programs prepare actual coded instructions that can be used by programmable manufacturing equipment, so one effect of the adoption of CAD may be to reduce the cost of adopting and using programmable manufacturing equipment. Since the prices of that equipment are also falling, the effects of the two price changes on the adoption of that equipment are mutually reinforcing. Of course, CAD also makes it cheaper for the firm to adopt a broader product line and to update its products more frequently. If the firm does so, than an indirect effect is to make it more profitable to switch to more flexible manufacturing equipment that is cheaper to change over. So, this indirect effect reinforces the direct effects of the changing input prices. With short production runs, the firm can economize on inventory costs (such as interest, storage, and obsolescence) by scheduling production in a way that is quickly responsive to customer demand. Such a scheduling strategy increases the profitability of technologies that enable quicker and more accurate order processing, such as modern data communications technologies. So, another indirect effect of falling CAD prices coincides with the falling price of data communication equipment. Thus, CAD equipment, flexible manufacturing technologies, shorter production runs, lower inventories, increased data communications, and more frequent product redesigns are complementary. However, the complementarities do not stop at the level of manufacturing, but extend to marketing, engineering, and organization.

The marketing side of the analysis involves two additional elements besides those already mentioned. First, more frequent setups lower the inventory necessary to support a unit of sales and thus also the marginal cost of output. This encourages lower prices. The second element arises because buyers value fast delivery. If most customers have good alternative sources of supply and only a few are "locked in," then the resulting relationship between delivery time and demand is convex. In that case, reducing or eliminating production delays makes it profitable to reduce other sources of delay as well. Then, computerized order processing and a fast means of delivery are complementary to the quick responsiveness of the modern factory to new orders.

On the engineering side, as product life cycles become much shorter than the life of the production equipment, it becomes increasingly important to account for the characteristics of existing equipment in designing new products. At the same time, the emergence of computer-aided design has made it less costly to modify initial designs, to estimate the cost of producing various designs with existing equipment, and to evaluate a broader range of potential designs. These changes have contributed to the growing popularity among U.S. firms of "design for manufacturability," in which products are developed by teams composed of designers, process engineers, and manufacturing managers (Robert Hayes, Steven Wheelwright, and Kim Clark, 1988), and the corresponding practice among Japanese automakers of providing preliminary specifications to suppliers who comment on the proposed design and supply drawings of parts (Asanuma, 1988b)—innovations in engineering organization that contribute to a more efficient use of existing production equipment and manufacturing know-how. Moreover, taking account of the limits and capabilities of production equipment in the design phase makes it easier to ensure that quality standards can be met, and so is complementary to a marketing strategy based on high quality.

The firm's problem in deciding whether to adopt any or all of these changes is marked by important *non-convexities*. These are first of the familiar sorts associated with indivisibilities: product line sizes are naturally integer-valued. A form of increasing returns also figures in the model, because the marginal impact of increasing the speed with which customers are served increases the service speed. Beyond these, however, the complementarities noted above can be a further source of non-convexities that are associated with the need to coordinate choices among several decision variables. For example, purchase of CAD/CAM technology makes it less costly for a firm to increase its frequency of product improvements, and more frequent product introductions raise the return to investments in CAD/CAM technology. Thus, it may be unprofitable for a firm to purchase a flexible CAD/CAM system without changing its marketing strategy, or to alter its marketing approach without adopting a flexible manufacturing system, and yet it may be highly profitable to do both together. (In contrast, if the value of a smooth concave function at some point in the interior of its domain cannot be increased by a small change in the value of any single variable, then the function achieves a global maximum at that point.) These non-convexities then explain why the successful adoption of modern manufacturing methods may not be a marginal decision.

It is natural to expect the characteristics of the modern manufacturing firm to be reflected in the way the firm is managed and the way it structures its relations with customers, employees, and suppliers. Exploiting such an extensive system of complementarities requires coordinated action between the traditionally separate functions of design, engineering, manufacturing, and marketing. Also, according to transaction costs theories, the increasing use of flexible, general purpose equipment in place of specialized, single purpose equipment ought to improve the investment incentives of independent suppliers (Oliver Williamson, 1986; Benjamin Klein, Robert Crawford, and Armen Alchian, 1978; Jean Tirole, 1986) and to reduce cost of the negotiating short-term contracts (Milgrom and Roberts, 1987) and so to favor short-term contracting with independent suppliers over alternatives like vertical integration or long-term contracting. The supplier relations that mark modern manufacturing firms—involving close coordination between the firm and its independently owned contractors and suppliers—appear to be consistent with these theories, and inconsistent with theories in which joint planning can only take place in integrated firms.

In this essay, we develop a theoretical model of the firm that allows us to explore many of the complementarities in modern manufacturing firms. The non-convexities inherent in our problem makes it inappropriate to use differential techniques to study the effects of changing parameters. Instead, we utilize purely algebraic (lattice-theoretic) methods first introduced by Donald M. Topkis (1978), which provide an exact formalization of the idea of groups of complementary activities. In problems with complementarities among the choice variables this approach easily handles both indivisibilities and non-concave maximands while allowing sharp comparative statics results. In particular, we give conditions under which the set of maximizers moves monotonically with changes in a (possibly multidimensional) parameter. Because these methods are quite straightforward and would seem to be of broad applicability in economics, but are not well known among economists, we describe them in some detail in Section I.

Our model and its basic analysis are provided in Section II. The firm in our model choices its price; the length of the product life cycle or frequency of product improvements (a surrogate for quality); its order-receipt, processing, and delivery technologies; various characteristics of its manufacturing and design technologies as reflected in its marginal cost of production and its costs of setups and new product development; its manufacturing plan, including the length of the production cycle (and, implicitly, its inventory and back-order levels); and aspects of its quality control policy, all with the aim of maximizing its expected profits. Using reasonable assumptions about the nature and equipment costs, we find that the complementarities in the system are pervasive. We use the firm's optimizing response to assumed trends in input prices (the falling costs of communication, computer-aided design, and flexible manufacturing) in the presence of these complementarities to explain both the clustering of characteristics and the trends in manufacturing.

In Section III, we turn our attention to the organizational problems associated with the new technologies. We summarize and review the predictions of the model in the concluding Section IV.

I. The Mathematics of Complementarities

Here we review some basic definitions and results in the mathematics of complementarities. The results permit us to make definite statements about the nature of the optimal solution to the firm's problem and how it depends on various parameters, even though the domain of the objective function may be non-convex (for example, some variables may be integer-valued) and the objective function itself may be non-concave, non-differentiable, and even discontinuous at some points. For additional developments and missing proofs, see Topkis.

We first introduce our notation. Let $x, x' \in \mathbf{R}^n$. We say that $x \geq x'$ if $x_i \geq x'_i$ for all i. Define $\max(x, x')$ to be the point in R^n whose ith component is $\max(x_i, x'_i)$, and $\min(x, x')$ to be the point whose ith component is $\min(x_i, x'_i)$. This notation is used below to define the two key notions of the theory. The first notion is that of a supermodular function, which is a function that exhibits complementarities among its arguments. The second is that of a sublattice of $\mathbf{R}^n$, a subset of $\mathbf{R}^n$ that is closed under the max and min operations and whose structure lets us characterize the set of optima of a supermodular function.

Definition 1: A function $f: \mathbf{R}^n \to \mathbf{R}$ is *supermodular* if for all $x, x' \in \mathbf{R}^n$,

$$(1) \quad f(x) + f(x') \leq f(\min(x, x')) + f(\max(x, x')).$$

The function f is *submodular* is $-f$ is supermodular.

Inequality (1) is clearly equivalent to

$$[f(x) - f(\min(x, x'))] + [f(x') - f(\min(x, x'))] \leq f(\max(x, x')) - f(\min(x, x')):$$

the sum of the changes in the function when several arguments are increased separately is less than the change resulting from increasing all the arguments together. The inequality is also equivalent to

$$f(\max(x, x')) - f(x') \geq f(x) - f(\min(x, x')):$$

increasing one or more variables raises the return to increasing other variables. These reformulations of the defining inequality make clear the sense in which the supermodularity of a function corresponds to complementarity among its arguments.

Note that any function of a single variable is trivially supermodular. This observation serves to resolve various questions about possible relationships between supermodularity and other concepts. However, even in a multidimensional context, supermodularity is distinct from, but related to, a number of more familiar notions. First, supermodularity has no necessary relation to the concavity or convexity of the function: consider $f(x_1, x_2) = x_1^a + x_2^b$, which is supermodular for all values of a and b but may be either concave or convex (or both or neither). Nor, in the context of production functions, does supermodularity carry implications for returns to scale. For example, the Cobb-Douglas functions $f(x_1, x_2) = x_1^a x_2^b$ may show increasing or decreasing returns to scale but are supermodular for all positive values of a and b. This is most easily checked using Theorem 2, below, which states that a smooth function f is supermodular if and only if $\partial^2 f/\partial x_i \partial x_j \geq 0$ for $i \neq j$. Thus, if f is supermodular and smooth, then the smooth supermodular function $-f$ shows weak cost complementarities as defined by William Baumol, John Panzar, and Robert Willig (1982, pp. 74–75). Even without smoothness, it is easily shown that a submodular function that is zero at the origin shows economies of scope as defined by Baumol et al. More generally, submodularity is related to, but distinct from, the notion of subadditivity that figures centrally in the study of cost functions.[3] For example, any function of a single variable is submodular, but obviously not all such functions are subadditive. Meanwhile, the functions on $[0,1] \times [0,1]$ given by $f(x_1, x_2) = 1 + x_1 + x_2 + \varepsilon x_1 x_2$ are submodular for $\varepsilon > 0$, supermodular for $\varepsilon \geq 0$, and subadditive for all ε sufficiently close to zero in absolute value.

Six theorems about supermodular functions are provided here. The first four together provide a relatively easy way to check whether a given function is supermodular. Theorem 5 indicates how, in a parameterized maximization problem, the maximizer changes with changing parameters, while Theorem 6 characterizes the set of maximizers of a supermodular function. It is Theorem 5 that makes our comparative statics exercises possible.

Let $x_{\setminus i}$ denote the vector x with the ith component removed and let $x_{\setminus ij}$ denote x with the ith and jth components removed. Let subscripts on f denote partial derivatives, for example, $f_i = \partial f/\partial x_i$, $f_{ij} = \partial^2 f/\partial x_i \partial x_j$.

THEOREM 1: *Suppose $f: \mathbf{R}^n \to \mathbf{R}$. If for all i, j, and $x_{\setminus ij}$, $f(x_i, x_j, x_{\setminus ij})$ is supermodular when regarded as a function of the arguments (x_i, x_j) only, then f is supermodular.*

THEOREM 2: *Let $I = [a_1, b_1] \times \cdots \times [a_n, b_n]$ be an interval in $\mathbf{R}^n$ with nonempty interior and suppose that $f: I \to \mathbf{R}$ is continuous and twice continuously differentiable on the interior of I. Then f is supermodular on I if and only if for all $i \neq j$, $f_{ij} \geq 0$.*

Theorem 2 is stated above in the form given in Topkis. For our application, we will need a slightly stronger theorem in which the condition that f is twice continuously differentiable is weakened to the condition that it can be written as an indefinite double integral with a nonnegative integrand. The precise extension is stated and proved in the Appendix.

THEOREM 3: *Suppose that f, $g: \mathbf{R}^n \to \mathbf{R}$ are supermodular functions. Then $f + g$ is supermodular. If, in addition, f and g are nonnegative and nondecreasing, then fg is supermodular.*

THEOREM 4: *Suppose that $f: \mathbf{R}^{1+n} \to \mathbf{R}$ is supermodular and continuous in its first argument. Then for all a, $b \in \mathbf{R}$, the function $g: \mathbf{R}^n \to \mathbf{R}$ defined by $g(x) = \max_{y \in [a,b]} f(y, x)$ is supermodular.*

PROOF:

Since f is continuous in its first argument, the function g is well defined. For all x and x', there exist y and y' with $g(x) = f(y, x)$ and $g(x') = f(y', x')$. Then,

$$\begin{aligned} g(x) + g(x') &= f(y, x) + f(y', x') \\ &\leq f(\max(y, y'), \max(x, x')) \\ &\quad + f(\min(y, y'), \min(x, x')) \\ &\leq g(\max(x, x')) \\ &\quad + g(\min(x, x')). \qquad \square \end{aligned}$$

[3] A function f is subadditive if $f(x) + f(y) \geq f(x + y)$ for all x and y.

In what follows we will be particularly concerned with constrained optimization of supermodular functions, and our results will depend on the constraint set having the right structure or shape, namely, that of a sublattice of $\mathbf{R}^n$.

Definition 2: A set T is a *sublattice* of $\mathbf{R}^n$ if for all $x, x' \in T$, $\min(x, x') \in T$ and $\max(x, x') \in T$.

In our application, the definition of a sublattice represents the idea that if it is possible to engage in high (respectively, low) levels of each of several activities separately, then it is possible to engage in equally high (resp., low) levels of all of the activities simultaneously. Thus, for example, if $S_1, \dots, S_n$ are arbitrary subsets of $\mathbf{R}$, then $S_1 \times \cdots \times S_n$ is a sublattice of $\mathbf{R}^n$. However, the product sets are not the only sublattices. The sublattice structure also permits the possibility that some activities can be engaged in at a high level *only if* the others are also carried out at a high level. For example, if $x \geq x'$, then $\{x, x'\}$ is a sublattice.

Definition 3: Given two sets $S, S' \subset \mathbf{R}^n$, we say that S is *higher than* S' and write $S \geq S'$ if for all $x \in S$ and $x' \in S'$, $\max(x, x') \in S$ and $\min(x, x') \in S'$.

THEOREM 5: *Suppose* $f\colon \mathbf{R}^{n+k} \to \mathbf{R}$ *is supermodular and suppose* $T(y)$ *and* $T(y')$ *are sublattices of* $\mathbf{R}^n$. *Let* $S(y) \equiv \operatorname{argmax}\{f(z, y) | z \in T(y)\}$, *and define* $S(y')$ *analogously. Then* $y \geq y'$ *and* $T(y) \geq T(y')$ *imply that* $S(y) \geq S(y')$.

PROOF:

Let $x \in S(y)$ and $x' \in S(y')$ and $y \geq y'$ so that $y = \max(y, y')$ and $y' = \min(y, y')$. Since $T(y) \geq T(y')$, $\max(x, x') \in T(y)$ and $\min(x, x') \in T(y')$. From the definitions, $f(x, y') \geq f(\max(x, x'), y)$ and $f(x', y') \geq f(\min(x, x'), y')$, but since f is supermodular, $f(x, y) + f(x', y') \leq f(\max(x, x'), y) + f(\min(x, x'), y')$ from which the conclusion is immediate. □

THEOREM 6: *Suppose* $f\colon \mathbf{R}^n \to \mathbf{R}$ *is supermodular and suppose* T *is a sublattice of* $\mathbf{R}^n$. *Then the set of maximizers of* f *over* T *is also a sublattice.*

PROOF:

Apply Theorem 5 with $y = y'$. □

Theorem 5 is particularly important for our application. When its conclusion holds, we shall say that the set of optimizers "rises" as the parameter values increases. What justifies this language? The theorem implies, for example, that if $x^*(y)$ and $x^*(y')$ are the unique maximizers given their respective parameter vectors y and y' and if $y \geq y'$, then $x^*(y) \geq x^*(y')$. (For uniqueness implies that $x^*(y) = \max(x^*(y), x^*(y'))$, from which $x^*(y) \geq x^*(y')$ follows.) Alternatively, suppose we assume that f is a *continuous* supermodular function that T is *compact* sublattice, so that the set of maximizers corresponding to any parameter vector y is compact. Then, by Theorem 5, there are greatest and least elements $\bar{x}(y)$ and $\underline{x}(y)$ in the set of maximizers S. One can show that both $\bar{x}(y)$ and $\underline{x}(y)$ are nondecreasing functions of y. (Using Theorem 6 and the definitions, $\bar{x}(y') \leq \max(\bar{x}(y), \bar{x}(y')) \leq \bar{x}(y)$ and similarly $\underline{x}(y) \geq \min(\underline{x}(y), \underline{x}(y')) \geq \underline{x}(y')$.)

II. Complementarities in Production

We study a model of a multiproduct firm facing a downward sloping demand curve. The firm may be a monopoly or monopolistic competitor. Alternatively, our model may be viewed as a building block for a model of oligopolistic markets.

In the formal model, the firm chooses the levels of the following decision variables:

Variable	Interpretation
p	Price of each product
q	(Expected) number of improvements per product per period
a	Order receipt and processing time
b	Delivery time
c	Direct marginal costs of production
d	Design cost per product improvement
e	Extra set-up costs on newly changed products

m	Number of setups per period
r	Probability of a defective batch
s	Direct cost of a setup
w	Wastage costs per setup

In addition, we denote the number of products by n.

The functional relationships and parameters that complete the model include the demand specification, the specification of the capital costs of different levels of the technological variables, the functional relation linking the average delay between receipt of an order and its being filled to the number of products and of setups, the marginal cost of production, the marginal cost of reworking defectives, the cost of holding inventories, and a time parameter that will proxy for the state of technology and demand. More specifically, we have

Parameter	Interpretation
ρ	Marginal cost of reworking a defective unit
τ	Calendar time
ι	Cost of holding inventory per unit
κ	Capital costs ($\kappa = \kappa(a, b, c, d, e, r, s, w, \tau)$)
μ	Base demand per product ($\mu = \mu(p, q, n, \tau)$)
δ	Demand shrinkage with delay time ($\delta = \delta(t, \tau)$)
ω	Expected wait for a processed order to be filled ($\omega = \omega(m, r, n)$)

The total expected wait for an order to be received, processed, filled, and shipped, which determines the value of the shrinkage factor (δ) on demand, is $t = a + \omega + b$, and realized demand is then $\mu(p, q, n, \tau)\delta(a + \omega + b, \tau)$. Thus, the firm's payoff function Π is

$$\begin{aligned}\Pi(p,q,m,a,b,c,d,e,r,s,w,\tau) &= (p - c - r\rho - \iota/m)n\mu(p,q,n,\tau)\\ &\quad \times \delta(a + \omega(m,r,n) + b, \tau)\\ &\quad - m(s + w) - nq(d + e)\\ &\quad - \kappa(a,b,c,d,e,r,s,w,\tau).\end{aligned}$$

The total profit Π is the operating profit minus the fixed costs associated with machine setups, product redesign, and the purchase of capital equipment.

The first term $(p - c - r\rho - \iota/m)n\mu\delta$ is the operating profit. For each unit sold, the firm receives the price p and pays direct production costs, expected rework costs, and inventory holding and handling costs. In line with the Economic Order Quantity models commonly used for inventory analysis, we treat the average levels of work-in-process and finished goods inventories as being directly proportional to demand and inversely proportional to the number of setups. Similarly, back orders are directly proportional to demand and decreasing in the number of setups. We have used the function δ, which we take to be uniform across products, to model the cost of back orders; this takes the form of lost demand when delivery is delayed. We have also modeled the firm as setting a uniform price (p) across the product line, which is reasonable given the symmetry of the products in the model.

The second term $m(s + w)$ is the cost of the setups, which, as suggested above, consists of the number of setups times the sum of the direct costs plus wastage per setup. The term $nq(d + e)$ is the cost of redesign over the period, including the extra setup costs on newly altered products. In this term, nq is the total number of redesigns or improvements.

The last term is κ, which is the capital cost of selecting the various technological variables, a, b, c, d, e, r, s, and w, at any date τ. Among these technology variables, the order receipt and processing time (a) is determined by the technology used for communicating orders (mail, express courier, FAX, electronic data communications networks, etc.) and by the means used to handle orders once received (manual entry, computerized order entry systems). Whichever choices are made, there are capital costs involved in setting up the corresponding systems. Similarly, different options exist that determine the speed of delivery (b) from inventory, and these too have differing capital costs.

Our model allows us to represent many aspects and tradeoffs in the firm's choice of manufacturing strategy. For example, the flexibility of design technology is modeled by the variable d. The introduction of computer aided design (CAD) lowers these marginal costs of redesigning and improving products, but it also involves significant capital expenditures on training, hardware, and software. Both of these effects are captured in our profit function.

Flexibility of manufacturing equipment has a number of aspects, several of which are represented in our model. First, flexibility is often associated with low costs of routinely changing over from producing one good to another. Here, this effect is represented first through the variable s: more flexible equipment means lower setup costs in terms of the downtime and direct labor costs involved in resetting the machines, switching dies, etc. Also, more flexible equipment might involve less wastage (lower w) per setup. This wastage might be in the form of extra inspection, scrap, rework, and repair costs that are necessary when a changeover is made. The precision of computer aided equipment, "design for manufacturability" (facilitated by CAD-CAM), and similar investments lower these costs. Finally, flexibility might involve costs of changing machinery over to produce new or redesigned products (low values of e).

The technological quality variable (r) captures a somewhat different feature of modern manufacturing methods. Improving quality on this dimension may involve investing in more precisely controlled machinery which may even constantly monitor and adjust itself. It may also improve more prosaic but possibly more significant efforts aimed at changing attitudes toward quality, such as giving workers the ability to stop the production line when a problem arises.

Although we do not explicitly model labor force decisions here, an element of the flexibility of modern manufacturing is associated with broadly trained workers and with work rules that facilitate frequent changes in activities. In this context we may interpret investments in flexibility in terms of worker education and industrial relations efforts, as well as the purchase of physical capital. Certainly, flexibility in the labor force and in the capital equipment are mutually complementary.

Finally, the choice of c, the marginal costs of production, has capital cost implications, if only through investing in learning how to control costs.

Even before we make any assumptions about the form of the unspecified functions κ, μ, δ, and ω, certain complementarities are evident in the model. For example, with more frequent changeovers (higher q), the returns to more efficient technologies for redesigning products and changing over equipment (higher values of $-d$ and $-e$) will naturally rise and, conversely, more efficient changeover and redesign technologies raise the marginal returns to increasing q. Similarly, an increase in the number of setups per period (m) and the concomitant reduction in inventories and back orders is complementary with a reduction in the components of set-up costs (increases in $-s$ and $-w$). Technologically, one expects that reduced set-up and changeover costs are bundled together in the new equipment. However, conclusions such as that one require that we make an assumption about the properties of the unspecified function κ. To make further statements, we need to make assumptions about properties of all the unspecified functions.

Our assumption about the form of δ is the following:

ASSUMPTION A1: *δ is twice continuously differentiable, nonnegative, decreasing and convex in t, nondecreasing in τ and submodular in (t, τ).*

The assumption that δ is decreasing in t simply means that increased delay reduces sales, while convexity says that the larger is the delay, the smaller is the marginal impact of additional delay. Inclusion of τ allows for a time trend in demand through δ. This trend must be nonnegative for our results, but it could be trivial. The submodularity assumption means that, as time passes, δ_t becomes weakly more negative or, equivalently, the returns to reducing waiting time, $-t$, increase weakly. This might come about because the adoption of more modern manufacturing methods by the firm's customers raises the importance of speedy service to them.

An immediate implication of the convexity assumption in Assumption A1 is that activities that reduce the several components of delay time (a, b, and ω) are mutually complementary, as can be easily verified by checking that the corresponding mixed partial derivatives of the profit function are positive. These complementarities may seem surprising, since the three components of waiting time are perfect substitutes for one another in determining the total delay. However, as our analysis shows, the possibility of substituting these elements to achieve a fixed time delay is irrelevant to their assessment as potential complements within the corporate strategy.

To complete our evaluation of the complementarities associated with speed, we must take an assumption about the ω function. Generally, we would expect ω to be increasing in the probability of a batch requiring reworking (r) and decreasing in the number of setups (m). The one nonobvious element in Assumption A2 below is that r and m are complements in determining ω: an increase in the number of setups (or decrease in batch size) is assumed to raise the impact on delay time of an increase in the probability of a batch being defective. This complementarity may be caused by the more frequent changeovers in the rework facility being required by more frequent changeovers in the main facility. Otherwise, it seems natural to expect the effect to be zero, which is also consistent with our assumption.

ASSUMPTION A2: *$\omega = \omega(m, r, n)$ is twice continuously differentiable, decreasing in m, increasing in r and supermodular in m and r, given n. That is, $\omega_m \leq 0$, $\omega_r \geq 0$, and $\omega_{mr} \geq 0$.*

As a consequence of (A1) and (A2), $-a$, $-b$, m and $-r$ are mutally complementary in increasing demand.

To complete our analysis of the marketing aspects of strategy, the form of the demand function μ must be restricted. We make two assumptions. The first is a standard one:

ASSUMPTION A3: *μ is twice continuously differentiable, increasing in q, and decreasing in p, while operating profits, defined by $(p - c - rp - \iota/m)n\mu(p, q, n, \tau)\delta(a + \omega + b, \tau)$, are a strictly quasi-concave function of p.*

The first part asserts innocuously that consumers prefer lower prices and higher quality. Since we will hold n fixed, we need make no assumptions on its effect, although it would be natural to assume that $n\mu$ is increasing in n. The assumption that demand is quasi-concave in prices is standard. The nonstandard part of our assumption about demand is contained in Assumption A4:

ASSUMPTION A4: *$\mu(p, q, n, \tau)$ is nondecreasing in τ and supermodular when regarded as a function of $-p$, q and τ, for given n.*

A4 is a complicated assumption. It would be satisfied, for example, by a multiplicatively separable specification of demand, $\mu = A(p)B(q)C(\tau)$, as well as by additively separable demand $A(p) + B(q) + C(\tau)$, and $A' < 0$, $B' \geq 0$, and $C' \geq 0$ in both cases. It asserts that the quantity demanded becomes (weakly) more sensitive to price and quality with passing time, and that at higher quality levels the quantity demanded is more sensitive to price changes. Again, we emphasize that we allow demand to be independent of τ, but if a dependence exists, it should not be such as to offset the supply-side effects of technological progress embodied in the effect of τ in the κ function.

Our final assumption is the following one, on the κ function. It is key because it embodies the presumed technological changes in the capital goods industries supplying the firm that are the basis for our arguments.

ASSUMPTION A5: *$\kappa(-a, -b, -c, -d, -e, -r, -s, -w, \tau)$ is submodular.*

This assumption is stated in terms of the negatives of the natural decision variables because a and b decrease with improved communication systems, better data transmission, entry, storage, manipulation, and retrieval systems, and speedier delivery methods, and the other choice variables decrease with improved design, manufacturing, quality, and cost-control technologies, that is, with increases in τ.

Conceptually, A5 has two parts. The first is our assumption about the time path of exogenous technological change: the incremental capital costs of modern technologies for communication, delivery, design, and flexible production are falling over time. Assuming differentiability of κ, so that Theorem 2 applies, these trends are captured in the inequalities $\partial^2\kappa/\partial x\,\partial\tau \leq 0$, $x = -a, -b, -c, -d, -e, -r, -s, -w$: technological change among capital equipment suppliers lowers the costs over time of the firm's increasing delivery speeds, using more flexible manufacturing methods, reducing the probability of defects, reducing costs of redesign and controlling production costs. Notice that we require no assumptions about the relative rates at which these prices are falling, because all these price changes will turn out to have mutually reinforcing effects.

The second part of A5 concerns the interrelationships among investments in the new technologies. For example, assuming that the mixed partial derivative of κ with respect to $-d$ and $-e$ is positive means that the level of investment in flexible equipment necessary to reduce extra set-up costs by a given amount is reduced by investments in flexible design equipment. Of course, if the technologies were completely separable, the condition would be met but, as we argued in the introduction, separability is not a realistic assumption. The cost of instituting both computer-aided design and flexible machining systems (FMS) to achieve given levels of design and set-up costs is generally less than the sum of the costs of instituting the two separately because the CAD equipment may provide set-up instructions readable by the FMS machinery, eliminating the costly step of encoding (possibly with error) the design instructions into a form readable by the flexible machine. Other complementarities in the physical equipment are similarly represented in κ. Thus, CAD makes it less costly to reduce defects by making it much cheaper to design products that are easily manufactured, while a computerized order-entry system can eliminate the need to transcribe order information into a form readable by the manufacturing computers, saving costs and reducing errors.

Other interactions, based on substitute uses of resources, could work against our complementary assumptions. For example, if the firm faces a fixed capital budget or a rising cost of capital with increased levels of investment, or if there are constraints on space or personnel and computer based systems for communication and design compete for these resources, then an investment in lowering a would raise the cost of investments to lower d. The second part of A5 is the hypothesis that the technological complementarities we have identified are larger than the effects of any constraints on the resources that the systems must share.

Note that throughout the analysis, we are holding the rework cost parameter ρ and the inventory holding cost parameter ι fixed.

The problem of maximizing Π is not amenable to standard, calculus-based techniques. First, although demand is assumed to be a quasi-concave in price, we have made no other concavity assumptions. Indeed, the assumed convexity of δ means that profit, exclusive of capital costs, is actually *convex* in the total delay between placing an order and receiving shipment, and since A5 places no restrictions on the concavity or convexity of κ, Π may well be convex in $-a$, $-b$, m and $-r$ over some ranges. In this case, satisfaction of a first-order condition identifies a (local) minimum with respect to the variable in question. Moreover, it is natural to take m to be integer-valued of the form, nk, where k is the number of production cycles per period and n is the number of products. However, the methods developed in the previous section are applicable here, once we have the necessary complementarities. In this, A5 plays a major role.

We are now ready to state and prove our main results. The idea is to use Theorem 2 to show that the firm's objective function is supermodular in the firm's (sign-adjusted) decision variables and that consequently, by Theorems 5 and 6, the set of optimizers forms a sublattice that moves up over time. However, it is not true that Π is a supermodular function of all its arguments, because the mixed partial derivatives of that function in price and the determinants of waiting time have the wrong sign when the

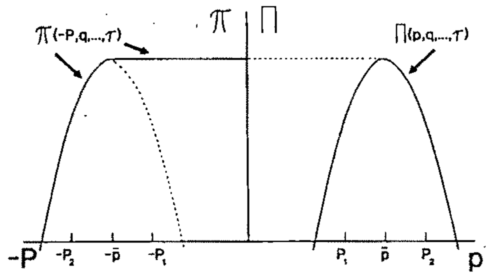


FIGURE 1. OBTAINING π FROM Π
NOTE THAT
$\max_{p \geq P_1}\Pi = \Pi(\bar{p},\ldots,\tau) = \pi(-\bar{p},\ldots,\tau)$
WHILE $\max_{p \geq P_2}\Pi(p,\ldots,\tau) = \Pi(P_2,\ldots,\tau)$
$= \pi(-\mathrm{P}_2,\ldots,\tau)$

price is set too low. To get around that difficulty, we consider the optimized value of profit with respect to price where the price is restricted by a lower bound P. (See Figure 1.) Letting this bound replace the price as the choice variable in our problem, it is apparent that this change of variables leaves the optimal values of the non-price variables unchanged. By A3, for any fixed values of the other decision variables and parameters, the corresponding optimal value of p is unique, and the *highest* optimal value of P equals the optimal price p.[4] Moreover, as we now show, the new function is super-modular in the sign-adjusted decision variables.

THEOREM 7: *Assume A1 through A5. Then the function*

$$\pi(-P,q,m,-a,-b,-c,-d,-e,-r,$$
$$-s,-w,\tau)$$
$$\equiv \max_{p \geq P}\Pi(p,q,m,a,b,$$
$$c,d,e,r,s,w,\tau)$$

is supermodular on the sublattice of $\mathbf{R}^n$ *defined by the restrictions that all the decision variables be nonnegative.*

PROOF:

By Theorem 3 and A5, it is enough that the function $\pi+\kappa$ be supermodular. If $\pi+\kappa$ were twice continuously differentiable everywhere, then by Theorem 2 it would be sufficient to check that all the cross-partials of $\pi+\kappa$ are nonnegative. The verification is routine, except that $\pi+\kappa$ may have no second derivatives on the set of points in the domain where $\partial(\Pi+\kappa)/\partial p = 0$. However, it is straightforward to verify that the slightly weaker assumptions of Theorem 2* in the Appendix are satisfied, so $\pi+\kappa$ is supermodular, as we had required. □

THEOREM 8: *Assume A1 through A5 and that κ is continuous. Let the individual decision variables each be constrained to lie in a compact set consistent with the nonnegativity requirement, so that together they lie in a compact set that is a sublattice. Then the set of maximizers of π is a compact sublattice which rises with τ.*

PROOF:

Apply Theorems 6 and 7. Note that this result allows us to restrict the decision variables to be integer-valued and to limit the number of available technologies to some finite set. The supermodularity of the functional form $\pi+\kappa$ is verified by investigating its derivatives on continuous intervals, and the restriction to a compact sublattice is imposed later in a way that permits a restriction to discrete choices.

The key conclusion is that the sign-adjusted decision variables all rise over time. Thus, as time passes, one expects to see a pattern of the following sort linking changes in a wide range of variables:

- Lower Prices,
- Lower Marginal Costs,
- More Frequent Product Redesigns and Improvements,
- Higher Quality in Production, Marked by Fewer Defects,
- Speedier Communication with Customers and Processing of Orders,
- More Frequent Setups and Smaller Batch Sizes, with Correspondingly Lower Levels of Finished-Goods and Work-In-Process

[4] This construct also ensures that the price at which demand is calculated will always weakly exceed marginal cost, $c + rp + \iota/m$.

Inventories and of Back Orders per Unit Demand,
- Speedier Delivery from Inventory,
- Lower Setup, Wastage, and Changeover Costs,
- Lower Marginal Costs of Product Redesign.

The conclusion in Theorem 6 that the set of optimizers forms a sublattice implies that if at any time there are multiple solutions to the optimization problem, then there is a highest and a lowest optimal solution in the vector inequality sense. Further, comparing any two firm's choices, if these differ, then the choice of selecting the higher, "more modern" level for each decision variable is also optimal, as is the vector made up of the term-by-term minimal values of the two firm's choices. More typically, however, we might expect a unique solution.

In any case, the chosen levels move up together over time in response to the falling costs of faster communications, more flexible production, and more frequent redesign.

The model we have presented is a static one, but it is nevertheless suggestive about the nature of the path to the modern manufacturing strategy. Specifically, it suggests that even if the changes that take place in the environment—especially the falling cost of the equipment used under the modern manufacturing strategy—happen gradually, the adoption process may be much more erratic, for two reasons. First, there are nonconvexities, which mean that the optimum may shift discontinuously, with the profit-maximizing levels of the whole complex of variables moving sharply upward. This makes it relatively unprofitable to be stuck with a mixture of highly flexible and highly specialized production equipment. One does not necessarily expect to find that the adoption of the new equipment is sudden; it may still be desirable to iron out the wrinkles in the new technology with an initial small scale adoption. What the theory suggests we should not see is an extended period of time during which there are substantial volumes of both highly flexible and highly specialized equipment being used side-by-side. Then, once the adoption is well underway, it should proceed rapidly, with increasing momentum.

Second, there are the complementarities, which make it relatively unprofitable to adopt only one part of the modern manufacturing strategy. The theory suggests that we should not see an extended period of time during which one component of the strategy is in place and the other components have barely begun to be put into place. For example, we should not see flexible equipment used for a long period with unchanging product lines.

The conclusion of Theorem 8 that firms will increase quality in the sense of reducing the probability r of a defective batch is worth further comment. Many observers have noted a focus on increased quality of output among modern manufacturing firms. One would expect that design for manufacturability would result directly in lower defect rates. However, the complementarities displayed in the model provide a second, less obvious incentive for increased quality. Decreases in the probability of defects are strictly complementary with increases in m through the effect on operating profit: demand grows with increases in m, and this increases the return to lowering costs by reducing the probability of reworking.[5]

Recall that we have held n, the number of products, fixed throughout this analysis. Inspection of the profit function in light of the arguments in Theorem 7 should make the necessity of doing this clear: neither n nor its negative are naturally complementary with the other decision variables. This shows up most clearly in the cost of redesign term, $-nq(d+e)$, where increases in n make decreases in d and e more attractive but increases in q less attractive. There are further potential complications through the demand term, and so without very special assumptions we cannot include n in the cluster of complements.

That we cannot include the number of products is somewhat surprising: surely broader product lines would seem to be

[5] Note too that decreases in r are also strictly complementary with increases in the other quality variable, q, as well as with decreases in the delay in communicating with customers and processing their orders (a) and in the time to deliver the inventory (b).

complementary with reduced set-up costs, and this intuition has in fact been verified in simpler models (see Xavier de Groote, 1988). However, the ambiguity surrounding n in richer models appear to reflect something real. On the one side, there are numerous examples of firms massively broadening their product lines with the adoption of modern manufacturing methods, and some of these were cited above. On the other, anecdotal evidence (for example, James B. Treece, 1989) as well as both the discussions of the "focused factory" found in the literature on manufacturing strategy (for example, David A. Garvin, 1988, especially Ch. 8) and some formal statistical analysis (Mikhel Tombak and Arnoud De Meyer, 1988) point to firms having reason to narrow their product lines when shifting to more modern manufacturing patterns and of their acting to do so.

III. Manufacturing and Organization

How is a manufacturing firm most efficiently organized and managed? Several of the trends analyzed in Section II have a direct bearing on this question. First, consider the complementarities that exist between the various functions in the firm: marketing, order-processing, shipping, engineering, and manufacturing. If the firm's problem were smooth and concave (despite the complementarities) and its environment were stationary and if the optimum is not on the boundary of the feasible set, the complementarities would not pose a serious organizational problem: if none of the managers controlling the individual functions can find a small change that raises the firm's expected profits, then there is no coordinated change —large or small—that can raise profits. However, in our non-concave problem, it is possible that only coordinated changes among all the variables will allow the firm to achieve its optimum. Non-convexities and significant complementarities provide a reason for explicit coordination between functions such as marketing and production.[6] (Extension of the methods in this paper to a game-theoretic context can be used to model this coordination problem and the role of the central coordinator: see Milgrom and Roberts, 1989.)

Even without non-convexities, significant complementarities in a rapidly changing environment provide another reason for close coordination between functions. Think of the managerial planning process as an algorithm to seek the maximum of the profit function. Successful performance in the face of rapid environmental change requires the use of fast algorithms (for example, Newton's method), and these require a coordinated choice of the decision variables that recognizes the interactions among these variables in the profit function.

Second, suppose that the organization being modeled is one where sales are made through several different stores. If the optimal speed of order-processing (a) jumps down, it may be desirable that all the stores install computerized systems linked to the manufacturing facility to track orders and sales. If there are fixed costs or other economies of scale in the computer system, then it is important that all, or nearly all, of the stores participate. However, unless all the costs and benefits of the change accrue to one agent, there arises a standard public goods, free-rider problem. Eliciting efficient cooperation from the store owners could be expensive and may provide a reason for vertical ownership of the distribution channel.

Third, Oliver Williamson and Klein, Crawford, and Alchian (1978) have argued that the advantages of increased vertical governance grow as assets become increasingly specialized. This occurs, it is argued, because the returns from specialized investments are vulnerable to appropriation. Then, as Williamson and Jean Tirole (1986) have argued, fear of appropriation causes insufficient investment to be made or, as we have argued (Milgrom and Roberts, 1987), it encourages the parties to waste resources by investing in bargaining position. Following this line of argument, let us equate "specialization" of assets with inflexibility of retooling to produce different products, so that it may be measured by e. The net costs of

[6] A similar point is made by de Groote, who investigates a different model of complementarities between marketing and manufacturing.

governance, bargaining, and deterred or distorted incentives are $\gamma(-v,-e)$, where v is a vector measure of the extent or complexity of vertical governance. We formalize a version of the hypothesis that increased flexibility of assets reduces the marginal value of governance activities with:

ASSUMPTION A6: *The function* $\gamma(-v, -e)$ *is sub-modular.*

THEOREM 9: *Assume that A1–A6 hold and consider the profit function*:

$$\pi(-p,m,q,-a,-b,-c,-d,-e,-r,$$
$$-s,-w,\tau)-\gamma(-v,-e).$$

Let each decision variable be constrained as in Theorem 8. Then the set of optimizers of $\pi-\gamma$ *is a sublattice and rises with* τ.

PROOF:

A direct consequence of Theorems 3, 5, 6, and 7, and A6. □

Thus, given Assumptions A1–A6 another predicted attribute in the characteristic cluster for flexible manufacturing companies is low vertical governance, for example, the extensive use of independently owned suppliers and subcontractors. This characteristic is an especially interesting one, given the usual conception of the difference between internal and market organization. Although uncertainty is not formally part of our model,[7] running this sort of "tight," low inventory operation with frequent redesigning of products in a world of uncertainties would surely require close coordination and communications with suppliers.[8] Yet according to our theory, the modern firm—despite its close relationships with suppliers and customers—will have little formal vertical governance..

Economists sometimes emphasize the need for close communication in the presence of supply or demand uncertainty as a reason for vertical integration (for example, Kenneth Arrow, 1975). If we were to formulate this alternative hypothesis using a submodular governance cost function $\lambda(m,v)$, we would arrive at the conclusion that v increases over time and that more extensive vertical governance is part of the cluster of characteristics of a modern manufacturing firm. The anecdotal evidence contained in press reports suggests to us that this conclusion is wrong, and that the former hypothesis A6 is the better one.

IV. Conclusion

The cluster of characteristics that are often found in manufacturing firms that are technologically advanced encompasses marketing, production, engineering, and organization variables. On the marketing side, these firms hold down prices while emphasizing high quality supported by frequent product improvements. Customers orders are filled increasingly quickly, with back-order levels being systematically reduced. In terms of technology, modern manufacturing firms exploit rapid mass data communications, production equipment with low setup, wastage, and retooling costs, flexible design technologies, product designs that use common inputs, very low levels of inventories (of both work in process and finished goods), and short production cycle times. They also seem to push differentially to increase manufacturing quality and. simultaneously, to control variable production costs. At the engineering and organizational levels, there is an integration of the product and process engineering functions and an extensive use of independently owned suppliers linked with the buying firm by close communications and joint planning.

We have argued in this paper that this clustering is no accident. Rather, it is a result of the adoption by profit-maximizing firms of a coherent business strategy that exploits complementarities, and the trend to adopt

[7] However, introducing uncertainty would cause no difficulties because the expectation of a supermodular function is supermodular. See Milgrom and Roberts, 1989.

[8] For a model of some aspects of this issue, see Milgrom and Roberts, 1988. In that model, inventories play a buffering role whose importance is reduced when communication is increased.

this strategy is the result of identifiable changes in technology and demand. Our formal model includes eleven decision variables from the claimed cluster of complements plus a parameter to account for the passage of time. There are thus 66 potential cross effects among the twelve variables, and all of these are nonnegative: there are extensive complementarities in marketing, manufacturing, engineering, design, and organization that make it profitable for a firm that adopts some of these characteristics to adopt more. We have also argued that the non-convexities in the problem mitigate against any smooth distribution of these characteristics among firms. For this reason, we are hopeful that empirical work will provide evidence of distinctly separated clusters of firm characteristics as support for our theory. Given our assumptions about time trends in prices, we also expect to find an increasing proportion of manufacturing firms adopting the modern manufacturing strategic cluster that we have described.

APPENDIX

THEOREM 2*: *Let $I = [a_1, b_1] \times \cdots \times [a_n, b_n]$ be an interval in $\mathbf{R}^n$ with nonempty interior and let $f: I \to \mathbf{R}$. Suppose that for every pair of arguments ij, there exists a function $f_{ij}: I \to \mathbf{R}$ such that f is the indefinite integral of f_{ij}. That is, for fixed $x_{\setminus ij}$ and for $x_i' > x_i$ and $x_j' > x_j$,*

$$f(x_i', x_j', x_{\setminus ij}) + f(x_i, x_j, x_{\setminus ij}) - f(x_i', x_j, x_{\setminus ij}) - f(x_i, x_j', x_{\setminus ij}) = \int_{x_i}^{x_i'} \int_{x_j}^{x_j'} f_{ij}(s, t, x_{\setminus ij})\, ds\, dt$$

If each f_{ij} is nonnegative, then f is supermodular on I.

Remark 1: *In our application, f is continuous on I and twice continuously differentiable on a set S with $\partial^2 f/\partial x_i \partial x_j \geq 0$ on S. Moreover, for all $\bar{\bar{x}}_{\setminus ij}$ the set $(I - S) \cap \{x | x_{\setminus ij} = \bar{\bar{x}}_{\setminus ij}\}$ is a curve. So, taking $f_{ij} = \partial^2 f/\partial x_i \partial x_j$ where defined and $f_{ij} = 0$ elsewhere, Theorem 2* implies that f is supermodular.*

PROOF:

In view of Theorem 1, it suffices to establish the conclusion for the case $n = 2$. Given any two unordered points x and x' with, say, $x_1 > x_1'$ and $x_2' > x_2$,

$$f(\max(x, x')) + f(\min(x, x')) - f(x) - f(x') = \int_{x_1'}^{x_1} \int_{x_2}^{x_2'} f_{12}(s, t)\, ds\, dt \geq 0,$$

from which it follows that $f(x) + f(x') \leq f(\max(x, x')) + f(\min(x, x'))$. □

REFERENCES

Abegglen, James and Stalk, George, Jr., *Kaisha: The Japanese Corporation*, New York: Basic Books, 1985.

Alster, Norm, "What Flexible Workers Can Do," *Fortune*, February 13, 1989, 62–66.

Arrow, Kenneth, "Vertical Integration and Communication," *Bell Journal of Economics*, Spring 1985, *6*, 173–83.

Asanuma, Banri, (1988a) "Manufacturer-Supplier Relationships in Japan and the Concept of Relation-Specific Skill," Kyoto University Economics Working Paper No. 2, 1988 (Forthcoming in the *Journal of the Japanese and International Economies.*)

______, (1988b) "Japanese Manufacturer-Supplier Relationships in International Perspective," Kyoto University Economics Working Paper No. 8, September 1988.

Baumol, William J., Panzar, John C. and Willig, Robert D., *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich, 1982.

de Groote, Xavier, "The Strategic Choice of Production Processes," unpublished doctoral dissertation, Stanford University, 1988.

Dumaine, Brian, "How Managers Can Succeed Through Speed," *Fortune*, February 13, 1989, 54–59.

Garvin, David A., *Managing Quality: The Strategic and Competitive Edge*, New York: Free Press, 1988.

Hayes, Robert H., Wheelwright, Steven C. and Clark, Kim B., *Dynamic Manufacturing: Creating the Learning Organization*, New York: Free Press, 1988.

Hausman, Warren, "Computer-Integrated Manufacturing: Lessons from Ten Plant Visits," Seminar presented at the Graduate School of Business, Stanford University, Stanford, CA, November 1988.

Henkoff, Ronald, "This Cat Is Acting Like a Tiger," *Fortune*, December 19, 1988, 69–76.

Hounshell, David A, *From the American System to Mass Production: 1800–1932*, Baltimore: Johns Hopkins University Press, 1984.

Kiechel, Walter III, "Corporate Strategy for the 1990s," *Fortune*, February 29, 1988, 34–42.

Klein, Benjamin, Crawford, Robert and Alchian, Armen, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *26*, 297–326.

Milgrom, Paul and Roberts, John, "Bargaining and Influence Costs and the Organization of Economic Activity," Discussion Paper, Graduate School of Business, Stanford University, 1987. (Forthcoming in J. Alt and K. Shepsle, eds., *Positive Perspectives on Political Economy*, Cambridge: Cambridge University Press.)

______ and ______, "Communication and Inventories as Substitutes in Organizing Production," *Scandinavian Journal of Economics*, 1988, *90*, no. 3, 275–289.

______ and ______, "Rationalizability, Learning and Equilibrium in Games with Strategic Complementarities," Discussion Paper, Graduate School of Business, Stanford University, 1989.

Mitchell, Russell, "Masters of Innovation: How 3M Keeps Its New Products Coming," *Business Week*, April 10, 1989, 58–63.

Moore, Thomas, "Make or Break Time for General Motors," *Fortune*, February 15, 1988, 32–50.

O'Rourke, Tracy, "A Case for CIM," Lecture delivered at the Conference on Manufacturing, Stanford University, Stanford, CA, May 1988.

Peters, Tom, "Hats Off to Benetton's Apparel Network," *Palo Alto Times-Tribune*, November 18, 1987, p. E1.

Piore, Michael J., "Corporate Reform in American Manufacturing and the Challenge to Economic Theory," mimeo., Massachusetts Institute of Technology, 1986.

______ and Sabel, Charles F., *The Second Industrial Divide: Prospects for Prosperity*, New York: Basic Books, 1984.

Port, Otis, (1989a) "Smart Factories: America's Turn?" *Business Week*, May 8, 1989, 142–48.

______, (1989b) "The Best-Engineered Part Is No Part at All," *Business Week*, May 8, 1989, 150.

Taylor, Alex III, "Why Fords Sell Like Big Macs," *Fortune*, November 21, 1988, 122–28.

Tirole, Jean, "Procurement and Renegotiation," *Journal of Political Economy*, April 1986, *94*, 235–59.

Tombak, Mikhel and De Meyer, Arnoud, "Flexibility and FMS: An Empirical Appraisal," *IEEE Transactions on Engineering Management*, May 1988, *35*, 101–107.

Topkis, Donald M., "Minimizing a Submodular Function on a Lattice," *Operations Research*, March-April 1978, *26*, 305–21.

Treece, James B., "GM's Bumpy Ride on the Long Road Back,"*Business Week*, February 13, 1989, 74–78.

Valery, Nicholas, "Factory of the Future: Survey," *The Economist*, May 30, 1987, 3–18.

Williamson, Oliver, *Economic Institutions of Capitalism*, New York: Free Press, 1986.

Wright, Kenneth and Bourn, David, *Manufacturing Intelligence*, Reading, MA: Addison Wesley, 1988.

General Motors Corporation, "First a Vision, Now the Payoff," *General Motors Public Interest Report 1988*, Detroit, 1988, 2–15.

U.S. Congress, Office of Technology Assessment, *Computerized Manufacturing Automation: Employment, Education and the Workplace*, Washington, 1984.

Malthusian Selection of Preferences

By INGEMAR HANSSON AND CHARLES STUART*

We study natural selection of preferences using a golden-age model with endogenous population. In equilibrium, all agents have preferences with maximum biological fitness given resource constraints and total population is the maximum the environment can sustain. Naturally selected agents follow the golden rule, acting as if they maximize the undiscounted sum of per-capita felicities of current and future generations. Selected preferences and hence work, saving, consumption, and population density vary predictably with environmental differences.

The preferences observed in living populations may be partly the result of biological selection that took place during pre-industrial times. Consider, for instance, life in a Malthusian environment, by which we mean an environment in which per-capita consumption drives population growth. Suppose the environment contains two families or "clans," all members of the first prone by heritage to indolence and immediate consumption and all members of the second with stronger, heritable tendencies to work and save. With greater labor and capital, the second clan may have greater sustainable per-capita consumption and hence greater population growth. As time passes, this clan would make up an ever-increasing part of the total population. Put differently, the population would asymptotically be characterized by the strong preference to work and save.

We expand upon this idea in the current paper, first developing a model of equilibrium demographics in a pre-industrial Malthusian environment, then characterizing naturally selected preferences as the preferences that maximize the sustainable number of offspring, finally illustrating how the analysis can be used to study aspects of economic development.[1] The specific focus is on selection of preferences for saving and labor supply. We stress that saving here is restricted to *inter*generational saving, meaning sacrifice of consumption by one generation that benefits later (younger) generations. The most important forms of intergenerational saving are doubtless in-kind transfers of human capital via bearing, support, upbringing, and education of children. Other forms include bequests and cash transfers from older individuals and government to younger individuals.[2]

*University of Lund and University of California at Santa Barbara, respectively. Thanks are due to a referee and to George Borjas, Ulla Larsen, Göran Skogh, Larry Summers, Lars-Gunnar Svensson, and Jack Votey. Research support from the Bank of Sweden Tercentenary Foundation is gratefully acknowledged.

[1] Equilibrium models of population have been used by economic demographers—see, for example, Ronald Lee (1987). In broad terms, the evolution of behavior and of preferences is the subject of modern evolutionary biology—see Edward O. Wilson (1980) for a survey. Formal analysis of genetic and cultural mechanisms for the evolution of what economists might call preferences has been conducted by Charles Lumsden and Wilson (1981). Robert Boyd and Peter Richerson (1985) focus on cultural evolution of preferences. Studies of economic behavior in a biological perspective include W. D. Hamilton (1964), John Maynard Smith (1982), Armen Alchian (1950), Gary Becker (1976), Jack Hirshleifer (1977), Paul Rubin and Chris Paul (1979), Howard Margolis (1982), and Robert Frank (1987).

[2] Intergenerational saving is fundamentally different from *intra*generational saving (Franco Modigliani and Richard Brumberg, 1954), which includes pension annuities and other wealth accumulated early in an individual's life cycle and decumulated later in the same individual's life cycle. Empirical evidence suggests that intergenerational saving makes up a substantial portion of total saving. For instance, Laurence Kotlikoff and Lawrence Summers (1981) estimate that intergenerational saving may be as high as 80 percent of total saving. Modigliani (1986), on the other hand, places the

I. Concern for Own and Future Generations' Consumption

A. *Very-Long-Run Equilibrium*

We use a paradigm of non-intermarrying clans that each consists of a sequence of overlapping generations. A generation is economically active for only one time period and only one generation is active in any period.[3] Each generation in a given clan has the same behavioral tendencies, or in economic terms the same preferences, but different clans may have different behavioral tendencies. This effectively treats preferences as stable in that they are passed down without mutation from generation to generation within a clan.[4]

We assume that the total population of all clans is bounded by what biologists call the *carrying capacity of the environment*. Carrying capacity captures the idea that there is a problem of the commons. To model the problem, we abstract from migration and treat the natural environment as a crowded public good that enters as an input in each clan's production function. We begin by characterizing the equilibrium population when carrying capacity is constant and then generalize to allow carrying capacity to change at an exogenously given rate of technological growth, g. Denote the total population of all clans in period t by N_t. Then by choice of units, the level of the crowded public good in each clan's production function (the "effective" level of the crowded factor) is $1/N_t$. We assume that per-capita output net of depreciation in a specific clan is a strictly concave, twice continuously differentiable, increasing function of the levels of the public good and of per-capita capital in the clan, k_t. This output is $q(1/N_t, k_t)$, where $q(\cdot)$ is assumed stationary and invariant across clans. Because we wish to focus on the size of total population and not on the effective amount of the crowded factor, we invert $1/N_t$ and work below with the function $f(\cdot)$ defined by $f(N_t, k_t) \equiv q(1/N_t, k_t)$. An increase in the total population of all clans increases crowding of the natural environment, which reduces per-capita output, so $f_N < 0$. We assume that output equals zero if either input equals zero. The marginal product of capital, f_k, is assumed to satisfy the Inada conditions $f_k \to \infty$ as $k \to 0$ and $f_k \to 0$ as $k \to \bar{k} < \infty$. We also impose "survivability restrictions" to ensure that total population in the model does not vanish. Specifically, we assume that for any positive given value of per-capita capital, per-capita output becomes infinite as total population goes to zero; formally, for $k > 0$, $f(N, k) \to \infty$ as $N \to 0$. (This restriction is stronger than needed and is weakened below.)

Because consumption is a biological necessity, we assume that greater per-capita consumption in a clan, c_t, leads to greater fitness.[5] We represent the necessity formally by writing the rate of population growth from period t to $t+1$ in a clan as a twice continuously differentiable, stationary function $p(c_t)$, with $-1 \leq p(0) < 0$, $p(c_t) > 0$ for some $c_t < \infty$, and partial derivatives $p_c > 0$ and $p_{cc} \leq 0$. An interpretation in the spirit of Malthus is that greater average consumption in a clan results in better nutrition and protection from environmental hazards, which raises the average number of children that survive to bear offspring.[6]

share at as low as 20 percent. Neither of these studies fully includes intergenerational saving in human capital, which may be substantial (Jacob Mincer, 1974; Arleen Leibowitz, 1974).

[3] We model each generation as a representative agent, abstracting from free riding within a generation.

[4] Passage of preferences by Mendelian genetics or by learning from parents are both compatible with the treatment here.

[5] The fitness of a trait may be defined as the relative rate of population growth of bearers given the selective pressures of the environment.

[6] If preferences are stable across generations, then naturally selected preferences must reflect the effects of environmental pressures over generations so selection prior to the demographic transitions associated with industrialization would partly explain preferences today. The Malthusian assumption that fitness rises in per-capita consumption is intended to reflect conditions during much of this long stretch of economic history. (To the extent that greater per-capita consumption con-

In any period, the active generation in a given clan splits per-capita output between consumption and investment. The split satisfies

$$(1)\quad f(N_t, k_t) = c_t + [k_{t+1}(1+p(c_t)) - k_t].$$

With the model specified, we characterize naturally selected *behavior*. We do so by studying a very-long-run equilibrium in which there is no tendency for the mix of behaviors in the population to change as a result of selection.[7] We treat this equilibrium as a steady state in which per-capita capital and consumption are identical across clans and constant and the total population of all clans is constant.[8] To understand how our assumptions about crowding of the natural environment can lead to constant total population, note that with sufficiently great total population, the level of effective natural environment in each clan's production function $(1/N_t)$ would be low enough to imply levels of output and hence consumption that would make population growth negative. Similarly, sufficiently low total population would by the survivability restriction lead to great enough per-capita output and hence consumption, for any given, positive level of per-capita capital, to imply positive growth of total population.

Three conditions determine the very-long-run equilibrium values of per-capita capital, per-capita consumption, and total population in the commons. First, because these three variables are constant in steady state, the constraint (1) becomes

$$(2)\qquad f(N,k) = c + kp(c),$$

where un-subscripted variables denote steady-state values. Equation (2) is the standard steady-state condition that constant per-capita capital requires investment of $kp(c)$ in order to compensate for population growth.

Second, no clan can have $p(c) > 0$ in steady state and all clans that do not vanish asymptotically must have precisely zero growth, so the equilibrium per-capita consumption of any clan is uniquely determined by

$$(3)\qquad p(c) = 0.$$

We refer to (3) as the equilibrium condition for a Malthusian population in an environment with constant carrying capacity: it requires per-capita consumption at the level that just keeps total population constant.

Third, because very-long-run equilibrium entails full selection of behavior, we require that selected behavior have maximum fitness. This requirement ensures that equilibrium is locally stable in that the equilibrium population cannot be invaded successfully by a small clan with other feasible behavior that would induce the invader to grow faster than the clans in the equilibrium.[9] The fit-

tributes to greater fitness in *developed* economies by improving nutrition and protection from the environment, the current analysis also characterizes a force of selection in developed economies.) Note that except during the past 100–150 years in developed countries, wealthier individuals, who are individuals with greater total consumption, have indeed tended to have more children (Becker, 1981; George Boyer, 1989; also Lee, 1987). More recently in developed countries, one sometimes observes negative correlations between wealth and fertility, although the overall tendency may not be strong. To the extent that population growth has been relatively flat in per-capita consumption in post-industrial times, little selection of the sort modeled here may have taken place since industrialization, which would mean that the (stable component of the) preferences observed in today's populations would largely be explained by pre-industrial selection. In any case, it is irrelevant given our desire to characterize the preferences of currently living agents that household fertility and income may be inversely related today in developed countries: if selection is slow, selection today predictably affects preferences only in the distant future.

[7] Traditionally, *long-run* refers to allocations that reflect full adjustment of factors and outputs to given tastes and technology. We use *very-long-run* to reflect full selection of preferences in addition to adjustment of factors and outputs.

[8] The total population of all clans might be written as the product of the number of clans and the average size of a clan. Although this product is uniquely determined in equilibrium, the number or average size of clans are not.

[9] Thus equilibrium behavior is an *evolutionarily stable strategy* in the sense of Maynard Smith (1982).

ness of a trait or the relative rate of population growth of bearers of the trait due to the selective pressures of the environment is fully indexed by the absolute rate of population growth of bearers, or, via $p(\cdot)$, by the per-capita consumption of bearers. The condition for maximum fitness is thus the condition for maximum induced per-capita consumption. To derive the condition, note that for a given size of total population, (2) determines a clan's per-capita consumption as a differentiable, concave function of its level of per-capita capital. Thus treat c as a function of k, differentiate (2) with respect to k, and set $dc/dk = 0$ to obtain

$$(4) \qquad f_k(N,k) = p(c),$$

which is the golden rule. The simple but important conclusion: in an environment in which per-capita consumption drives population growth, selection implies a tendency toward golden-rule saving.

Conditions (2)–(4) determine unique very-long-run equilibrium values of per-capita capital, per-capita consumption, and total population in the commons. In equilibrium, all agents behave identically and total population is the maximum the environment can sustain in steady state (carrying capacity). To see that total population equals carrying capacity, note first that (3) yields the steady-state level of per-capita consumption consistent with zero population growth. Substitute this zero-growth consumption level into the steady-state condition (2), obtaining an equilibrium condition in N and k. By taking k as given and solving the condition for N, one can view the condition as determining the total population supported by the environment in steady state as a function of per-capita capital. Differentiate the expression with respect to k and set $dN/dk = 0$ to characterize the behavior that maximizes total sustainable population. The result is (4). Thus maximization of fitness implies maximization of total sustainable population here.[10]

We denote equilibrium values with stars and illustrate the relationships between k^*, c^*, and N^* in Figure 1. The figure corresponds to a numerical example in which $p(c) = c^{0.5} - 1$ and $f(N,k) = (k/N)^{0.5} - 0.5k + 0.5$. The solid peaked curve indicates the relationship between per-capita capital and the population growth rate (fitness) in a clan when the total population of all clans equals one. The equation for this curve is derived generally, which is to say for an arbitrary level of total population, by using the steady-state condition (2) and the population growth function to find the level of steady-state consumption and hence the rate of population growth consistent with a given level of per-capita capital. The broken curve is derived from the same equation except that total population is lower ($N = 0.96$) and hence steady-state per-capita consumption and population growth are greater for each value of steady-state capital. At a sufficiently low population level, the survivability requirement ensures that $f(N,k) > 0$. This occurs on the broken curve at points above zero. Such a low level of total population cannot be part of a very-long-run equilibrium, however, because clans that choose k at or close to the golden rule given by (4) would have $p(c) > 0$ and hence would always grow at a positive rate even though the capacity of the environment to support population is limited and constant. Thus if $N =$

[10] This means that equilibrium in our model is one of pure "K-selection," defined as selection that favors efficient use of resources at population levels at or near the maximum sustainable level (Wilson, 1980, pp. 39–49). Note that "K-selection" does not necessarily favor large numbers of offspring per individual. Selection that favors large numbers of offspring per individual is referred to as "r-selection," which typically describes evolutionary pressures in shifting habitats in which only a small share of total offspring find resources needed to survive and reproduce. In man, r-selection might characterize some aspects of a "frontier environment" in which total population is far below the maximum sustainable level. The equilibrium in our model could be converted to one of (deterministic) r-selection by removing feedbacks whereby increases in population are self-limiting; formally, this could be done by deleting total population from the model and by deleting the Malthusian condition (3). Equilibrium would then be characterized by two equations, (2) and (4), in two unknowns, k and c.

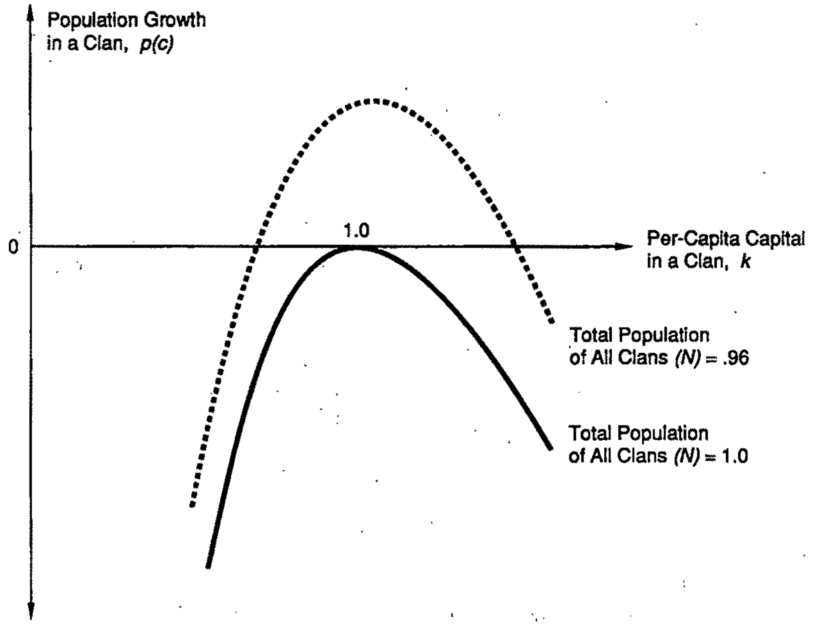


FIGURE 1

0.96 in the example, total population would tend to rise. The unique very-long-run equilibrium occurs in the example at $N^* = 1$ (solid curve) and $k^* = 1$. That is, because $p(c^*) = 0$ in equilibrium, we have $c^* = 1$ in this case. The steady-state condition is then $f = c^*$ or $(k^*/N^*)^{0.5} - 0.5k^* + 0.5 = 1$, and the golden rule is $f_k = p(c^*)$ or $0.5(k^*N^*)^{-0.5} - 0.5 = 0$, which together imply $N^* = 1$ and $k^* = 1$. Because no greater value of N is possible in steady state, total population equals the carrying capacity of the environment.

Very-long-run equilibrium may be compared to the traditional notion of long-run competitive equilibrium. Traditional long-run equilibrium is characterized by profit maximization and adjustment of the population of firms so that equilibrium profits equal zero. Here, very-long-run equilibrium in an environment without growth is characterized by population-growth (fitness) maximization and adjustment of total population so that equilibrium population growth equals zero. As traditional long-run equilibrium suggests the direction of market adjustment at any point of time, one may interpret very-long-run equilibrium as suggesting the direction of selection of behavior at a point of time.

It is straightforward to generalize the analysis to situations in which the carrying capacity of the environment grows at rates other than zero. Specifically, we denote growth-deflated population, $N_t/(1+g)^t$, by M_t and we take per-capita output in a given clan to be $f(M_t, k_t)$.[11] Writing production as a function of growth-deflated population in this way captures the idea that population is limited by carrying capacity in that, in steady state, per-capita output and capital are constant so growth-deflated population, $M_t = N_t/(1+g)^t$, must be constant, which is to say that total population must grow at the rate of increase in the carrying capacity of the environment. We continue to assume that crowding reduces per-capita output, or $f_M < 0$, that output equals zero if either input equals zero, and that $f(\cdot)$ satisfies the Inada conditions $f_k \to \infty$ as $k \to 0$ and $f_k \to 0$ as $k \to \bar{k} < \infty$. To ensure that total population

[11] That is, we assume that growth is purely "natural-environment-augmenting."

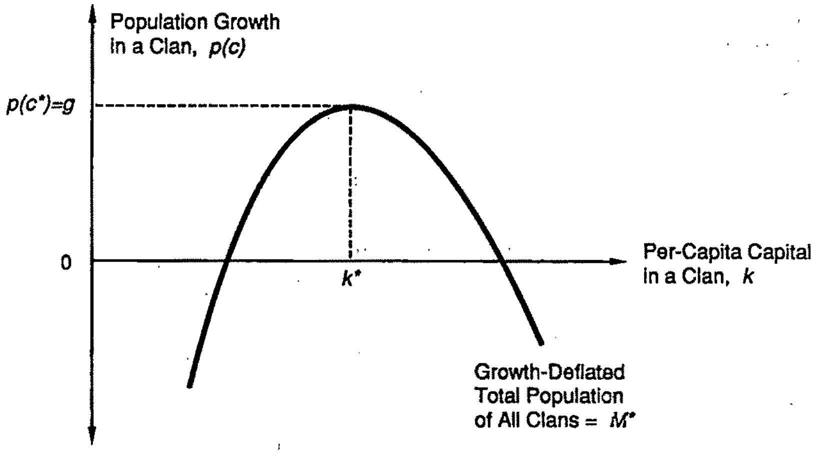


FIGURE 2

not vanish asymptotically, we assume that the level of consumption defined by $p(c^*) = g$ is finite and technologically feasible in that there are values $M>0$ and $k>0$ such that $f(M,k)-c^*-kg=0$.

To characterize the behavior selected in the very-long-run in nature, the steady-state resource constraint is now

$$(5)\qquad f(M,k)=c+kp(c).$$

Further, because the carrying capacity of the environment now grows at rate g, all clans that do not vanish asymptotically must grow at rate g. Thus per-capita steady-state consumption is uniquely determined by

$$(6)\qquad p(c)=g,$$

instead of by the earlier condition that $p(c)=0$. An implication of (6) is that per-capita consumption in steady state rises with the rate of technological growth. Thus if one defines subsistence consumption as the level of per-capita consumption that would leave the population of a clan unchanged, which is the root of $p(\bar{c})=0$, then positive technological growth would cause equilibrium consumption to exceed subsistence consumption.

Equation (4), which specifies that maximum fitness is achieved under the golden rule, continues to hold when g is not constrained to equal zero. Thus conditions (4)–(6) now determine unique very-long-run equilibrium values of capital, consumption, and growth-deflated total population. The equilibrium is illustrated in Figure 2, which is similar to Figure 1 except that the peak of the steady-state relationship between capital and the implied level of population growth has height g. Here, very-long-run equilibrium is characterized by population-growth maximization and adjustment of total population so that total population equals the carrying capacity of the environment at each point of time. (Thus the rate of population growth equals the rate of increase in carrying capacity.)

B. *Representing Selected Behavior with Preferences*

The very-long-run equilibrium characterized above describes naturally selected behavior. Economists generally represent behavior as the outcome of constrained optimization. Accordingly, we now represent naturally selected behavior as the outcome of constrained optimization and we characterize the preferences that are consistent with very-long-run equilibrium. There are many possible representations or "specifications." Although behavior in very-long-run equilibrium can be captured with simpler specifications, it is nonetheless useful to adhere to a traditional formulation in which generation z selects a consumption sequence that maximizes the discounted sum of the clan's current and future felicities to an infinite time horizon. One may think of this specification as a purely intergenerational rendering of Robert Barro's (1974) dynasty model.[12] The time horizon is assumed infinite because very-long-run equilibrium is a steady state. It turns out that the specific selection process posited here imposes restrictions on the felicity discount rate but does not uniquely determine the shape of the felicity function. In searching for the population-growth-maximizing discount rate, we therefore do not wish to restrict the discount rate to a range for which the discounted sum of an infinite sequence of felicities is finite. We thus assume a constant planning horizon of T periods, which is the same for all generations, and we interpret behavior over an infinite horizon as behavior when T goes to infinity.

In more detail, the contribution to generation z's utility from consumption in period $t > z$ is the discounted felicity of per-capita consumption in t times the ratio of the population in t to that in z,

$$(7)\quad u(c_t)[1+p(c_z)][1+p(c_{z+1})] \cdots [1+p(c_{t-1})]/(1+\delta)^{t-z},$$

where felicity is increasing, concave, and twice continuously differentiable with $u_c > 0$ and $u_{cc} \leq 0$, and where δ is the rate at which a generation discounts the felicity of consumption of the clan's future generations. A higher discount rate δ means that the current generation is more selfish or, prosaically, that parents show less love for progeny. It should be stressed that the analysis here is *inter*generational and not *intra*-generational, so δ is the geometric rate of *generation* preference and not the rate at which an individual might discount own future consumption.

To see the restrictions selection imposes on utility, we first characterize a clan's behavior given a utility function with arbitrary felicity and an arbitrary discount rate. For a given finite time-horizon, the first-best allocation for generation z maximizes

$$(8)\quad U_z = u(c_z) + u(c_{z+1}) \times \left(\frac{1+p(c_z)}{1+\delta}\right) + \cdots = \sum_{t=z}^{z+T} u(c_t) \prod_{v=z}^{t-1} \left(\frac{1+p(c_v)}{1+\delta}\right),$$

subject to budget constraints (1) and an initial value of k_z.[13] The first-order conditions that characterize the clan's behavior are

$$(9)\quad \left\{\frac{\partial u}{\partial c_t} + \frac{\partial p}{\partial c_t}\left[\sum_{v=t+1}^{z+T} u(c_v) \prod_{w=t+1}^{v-1} \left(\frac{1+p(c_w)}{1+\delta}\right)\right] \Big/ (1+\delta)\right\} \times \left[\prod_{v=z}^{t-1}\left(\frac{1+p(c_v)}{1+\delta}\right)\right] - \lambda\left(1 + k_{t+1}\frac{\partial p}{\partial c_t}\right) = 0,$$
$$z \leq t \leq z+T,$$

$$(10)\quad \lambda_t(f_k+1) - \lambda_{t-1}(1+p(c_{t-1})) = 0,$$
$$z+1 \leq t \leq z+T,$$

where the λ_t are Lagrange multipliers.

To represent selected behavior in the very-long-run, we find the preference parameter δ that solves (9) and (10) in very-long-run equilibrium when $T \to \infty$. Because each generation would want the same conditions to be satisfied in steady state, the equilibrium is time-consistent when the generations

[12] Andrew Abel and Kotlikoff (1988) provide empirical support for the Barro model.

[13] We define $\Pi_z^{z-1}\{\cdot\} \equiv 1$.

in the clan act sequentially. To proceed, take (9) for two adjacent periods, apply (10), and evaluate the resulting expression at $k_t = k^*$, $c_t = c^*$, and $M_t = M^*$ for all t:

$$(11)\quad \frac{\partial u}{\partial c}+\frac{\partial p}{\partial c}\left[\sum_{v=t+1}^{z+T} u(c^*)\left(\frac{1+p(c^*)}{1+\delta}\right)^{v-t-1}\right]\Big/(1+\delta) = \left[\frac{1+f_k}{1+\delta}\right]\left\{\frac{\partial u}{\partial c}+\frac{\partial p}{\partial c} \times\left[\sum_{v=t+2}^{z+T} u(c^*)\left(\frac{1+p(c^*)}{1+\delta}\right)^{v-t-2}\right]\Big/(1+\delta)\right\}.$$

It is easy to show that the discount rate that characterizes equilibrium behavior cannot be less than $p(c^*)$. To see this, suppose to the contrary that $\delta < p(c^*)$. Then as $T \to \infty$, (11) can only hold if $f_k \to \infty$, which requires that $k \to 0$. Because zero capital does not characterize fitness-maximizing behavior in very-long-run equilibrium, it must be that $\delta \geq p(c^*)$. In this case when $T \to \infty$, (11) implies

$$(12)\qquad \delta = f_k(M^*, k^*).$$

Equation (12) is based solely on the assumption that one can represent the behavior of a clan with positive steady-state capital as the outcome of maximizing (8). Under this representation, (12) says that one views the clan as accumulating capital to the point where capital's marginal product equals the clan's intergenerational discount rate. Consequently, if one observes a clan in a steady state with $k = k^* > 0$ and $M = M^*$, then the implied (or "indirectly observed") value of the discount rate must be precisely the marginal product of capital evaluated at the equilibrium values of total population and per-capita capital. Because a very-long-run equilibrium is a steady state, (12) also characterizes the naturally selected discount rate in very-long-run equilibrium when clans are seen as maximizing (8). Combining (12) with the golden rule (4), which is the condition for maximum fitness, we have the result that the naturally selected value of the discount rate equals the rate of population growth in very-long-run equilibrium:

$$(13)\qquad \delta^* = p(c^*).$$

That is, discounting by δ^* just balances the increasing weights that population growth places on the felicities of successive future per-capita consumption levels, so U_z is the simple sum $u(c_z) + \cdots + u(c_{z+T})$ to a distant horizon T. The intuition is that maximum fitness requires maximum steady-state per-capita consumption, which in turn requires that no generation be weighted less than any other generation in the utility of the economically active generation.[14] Clearly, this reflects substantial concern about the consumption of future generations.

C. *Implications and Discussion*

First and foremost, the analysis predicts that agents sacrifice substantial own consumption to provide for consumption of progeny. The prediction is consistent with observed sacrifices in the form of bearing, support, upbringing, and education of children, other transfers from older to younger individuals, and bequests. Such agreement between prediction and observation suggests that the analysis captures some of the forces that have shaped naturally selected preferences.

More precisely and more usefully, the analysis predicts that the value of the pure rate of generation preference (δ) can be ascertained directly as the rate of population growth over a long past. To illustrate, consider the European countries, which have relatively better historical demographic data than do the LDCs and which have had relatively less migration than developed former European colonies such as the United States. (The analysis above abstracts from migra-

[14] With $\delta > p(c^*)$, a suboptimality would arise as agents would attempt to shift resources to themselves from future generations by saving too little or by instituting pay-as-you-go Social Security (Hansson and Stuart, 1989).

tion, but one might as an approximation extrapolate values of δ from European countries to the United States because many U.S. immigrants came from Europe.) Good demographic data before about 1500 are scant for any country; a baseline might be Ansley Coale's (1974) estimate that the *world's* population grew at an estimated annual rate of 0.056 percent from the time of Christ to 1750. Population growth rates have been greater, however, in Europe in recent centuries. For a sampling, the yearly growth rate in Germany from 1816 to 1910 based on census data was 1.02 percent (John Knodel, 1974). German population growth was restrained by net emigration, but during 1871–1910, a period of relatively great emigration, "external migration reduced the population growth that would have accrued from natural increase alone by only 10 percent" (Knodel, p. 192). In England, yearly population growth occurred at a 0.35 percent rate between 1541 and 1751 and at a 1.11 percent rate between 1751 and 1871; net emigration rates over the period were generally less than 0.2 percent/year (E. A. Wrigley and R. S. Schofield, 1981). Population growth rates in France and Sweden from the mid-1700s to 1871 were typically somewhat lower than in England (Wrigley and Schofield, pp. 213–14). Thus the predicted value of δ in living Western populations might lie between perhaps zero and a few percent per year, or between perhaps zero and 1.0 per generation.

To derive further implications, one may think of the marginal product of capital in the clan's production function (f_k) as its real interest rate. Thus because (4) characterizes maximum fitness in any steady state given only that fitness increases in per-capita consumption and because (12) holds in any steady state in which behavior is represented as the outcome of maximizing (8), the analysis suggests that selection may act generally to drive real interest rates to equality with the rate of population growth.[15] Thus in steady state, the analysis views the real interest rate as based on productivity and not on time preference, because selection causes time preference to conform to productivity. Still more fundamentally, real interest rates today may have been determined by the growth of carrying capacity in pre-industrial times, because population growth conforms to the growth in carrying capacity in very-long-run Malthusian equilibrium and because this equilibrium is intended to characterize current preferences as the preferences selected over a relatively long past that includes pre-industrial times.

The analysis thus permits a simple prediction about real interest rates in modern times. From (12), the real interest rate equals the discount rate δ in equilibrium. We argued above that δ may lie between zero and a few percent per year. Today's real interest rate should therefore also lie between zero and a few percent. Modern data on real interest rates do not reject this prediction (see, for example, John Huizinga and Frederic Mishkin, 1983), but we emphasize that the current analysis neglects a large number of factors that could potentially influence the prediction. For instance, the analysis here abstracts from the effects of government policy and international trade.

The essence of the story of selection in this section is the hypothesis that among different saving behaviors (preferences), fitness is greatest for the saving behavior that leads to the greatest sustained level of per-capita consumption. That the predictions of the analysis conform roughly to observed data suggests that there is predictive value to the hypothesis. As a first analysis, however, our formalization abstracts from several forces of natural selection that may have influenced the preferences of currently living agents. One important concern is the puzzle that, contrary to Malthus's famous prediction, rates of population growth have persistently been lower than rates of growth in carrying capacity in industrialized economies. Apparently, demographic transitions in Western countries have involved "voluntary" reductions in the number of children per household. It is interesting to speculate on how natural selection could lead to vol-

[15] That is, equality of interest and population growth rates does not require the Malthusian condition (6) that population growth equal the growth in carrying capacity.

untary reductions in fertility at all, of course, because maximum relative population growth of a trait (fitness) would seem to require that carriers of the trait reproduce to the maximum sustainable extent allowed by available resources. Here is a good puzzle for future research.[16]

A second set of concerns is that some might argue that humans do not exhibit care for progeny to the extent that behavior is best described as maximization of the undiscounted sum of per-capita felicities to an unbounded horizon; that is, the current analysis may predict too little effective discounting of future generations. Note, however, that the current treatment abstracts from several real-world forces that might raise effective discounting of future generations, such as free riding within clans, random hazards in the environment, and expenditure of resources on war and defense. (In our analysis, interactions among clans are limited to negative "crowding" externalities in production.)

A third concern is that we do not undertake rigorous, global stability analysis. To do so carefully would involve the complication of specifying dynamics out of equilibrium when the population consists of coexisting clans that grow at different rates. It is easy to show, however, that if all clans were to set the same level of capital, total population would eventually converge to some level, and if all clans set $k=k^*$, growth-deflated total population would converge to M^*. Similarly, if $M=M^*$, clans that maintain constant per-capita capital of k^* would grow faster than clans that maintain any other levels of per-capita capital, so non-k^* clans would be outcompeted asymptotically by clans that set $k=k^*$.

A fourth concern is that there are many alternatives to (8), which specifies preferences as separable and Barro-dynastic. Nothing assures that the specification (8) of the form of concern for future consumption is correct. Alternatively, we could have specified generation z's optimization as one of maximizing $U(c_z, k_z)$ with $U=-\infty$ if $k_z \neq \alpha$ and $U>-\infty$ if $k_z=\alpha$. Malthusian selection of the parameter α would lead to $\alpha=k^*$ in very-long-run equilibrium so generation z would always leave a bequest of k^*. This alternative specification would predict the same tendency to save but would admit less flexibility of behavior than (8).[17]

A fifth concern is that the current analysis, like much of the work in evolutionary game theory, is "ecological." In particular, we assume that there is no mutation, no intermarriage, that no biological constraints restrict a clan's discount rate, and that, before very-long-run adjustments take place, there are many clans and a nondegenerate distribution of discount rates that includes δ^*. Perhaps the most important issue with the ecological approach here involves the assumptions that preferences are stable across generations and free of mutation. These assumptions may not describe selection of behavior by "cultural" processes involving innovation and imitation of other clans, which may be important in industrialized economies.[18]

[16] Analysts who take preferences as given have explained "voluntary" reductions in fertility as a response to greater costs of having children (Mincer, 1963; Robert Willis, 1973) or as related to increases in child quality (Becker and H. Gregg Lewis, 1973). There is doubtless truth in both explanations, but neither deals with the deeper issue of how voluntary reductions could be consistent with natural selection of preferences.

[17] Thus the current analysis does not predict the flexibility or sensitivity of saving behavior to changes in economic conditions. To predict this sensitivity, it would be necessary to characterize naturally selected behavior in an environment in which economic conditions change. A related point is that even if population growth is assumed strictly concave in per-capita consumption, the selection process here does not restrict felicity to be strictly concave in c or "risk averse." This is because we abstract from randomness in the environment. With such randomness, selection may determine the curvature of felicity.

[18] Such cultural selection may operate on time-frames of less than a single generation. Because we wish to study first the case of stable preferences, which is the traditional assumption of choice among economists, and because elements of preferences selected by rapid cultural processes may be relatively less stable biologically, we have abstracted from rapid cultural selection here. On the other hand, if agents imitate others who have greater per-capita consumption, then the results of the current analysis could be strengthened in that convergence to very-long-run equilibrium could be faster.

Before extending and applying the analysis in the next section, we note that the results to this point illustrate a fundamental property of naturally selected preferences: in equilibrium, marginal rates of substitution between variables that affect fitness equal marginal rates of substitution between variables in utility. A general demonstration of this property is in Section III. In the case of intergenerational saving, impacts on a clan's fitness may be taken as impacts on the clan's population in a future period. The population in future period x is

$$\prod_{v=z}^{x-1}[1+p(c_v)] \text{ times the population in period } z.$$

For $x > t+1$, the ratio of the derivatives of population with respect to c_t and c_{t+1}, which is the marginal rate of substitution between c_t and c_{t+1} as a determinant of population in period x, equals one in equilibrium because $c_t = c_{t+1}$. From (13), the marginal rate of substitution between c_t and c_{t+1} in the behavioral utility function, $(\partial U_z/\partial c_t)/(\partial U_z/\partial c_{t+1})$, also equals one in equilibrium.

II. Preferences to Work and Save, and an Application to Development Economics

A. *Overview*

We now extend the selection model to include a consumption-leisure tradeoff and we illustrate the usefulness of the approach by applying it to study the effects of the environment on very-long-run equilibrium. The analysis bears on Toynbee's (1956) "challenge and response" hypothesis, which explains economic development as a response to environmental challenge. Toynbee lists a number of historical examples of economic development in response to challenges of "hard countries," "new ground," "blows," (external) "pressures," and "penalizations," which include phenomena such as discrimination. The generality of Toynbee's hypothesis is controversial but can be studied using tools developed here. We consider a specific case, namely, that measured (material) economic development tends to rise and population density tends to fall with distance from the equator. Quite simply, the "challenge" in this case may be that surviving harsh winters requires shelter, clothing, secure storage facilities, and other material commodities not needed in the tropics. We capture this type of environmental challenge or "harshness" as a shift parameter in the fitness function $p(\cdot)$ and show that increases in it may lead to lower population density and greater per-capita labor, capital, and consumption in very-long-run equilibrium. Thus aspects of economic development may be explained by biological selection of economic behavior.

B. *Extending the Analysis*

Nature is probably too economical to give us a strong taste for something that does not raise fitness. In thinking about man's taste for leisure, for instance, one might argue that leisure contributes to fitness by providing valuable regeneration of the body, defense, exercise, education, social coordination, and conservation of calories. We summarize these benefits of leisure by including per-capita leisure (l) as an argument in the clan's population growth function with $p_l > 0$. We also include an environmental shift parameter, β (for bad), such that in a neighborhood of an initial very-long-run equilibrium, increased β corresponds to a harsher environment and means that population growth is lower, $p_\beta < 0$. Because output net of capital formation can be "consumed" as shelter, clothing, and storage, we assume that greater β increases the marginal fitness of consumption relative to that of leisure. To capture this simply, we assume that $p_{c\beta} > 0$ and $p_{l\beta} \leq 0$. We include per-capita labor, $h_t \equiv 1 - l_t$, as an additional argument in production and abstract from influences of the environment on production so per-capita output in a clan is $f(M_t, k_t, h_t)$. We maintain our earlier assumptions about $f(\cdot)$ and assume that an increase in any input, including the effective among of natural environment, raises the marginal products of the other two inputs. Because a rise in population increases crowding of the natural environment, which reduces f_k and f_h, we have $f_{Mk} < 0$ and $f_{Mh} < 0$. To allow behavior to adjust via selection to environmental conditions, we include l_t and β as arguments in $u(\cdot)$.

Once again we wish to characterize the behavior that maximizes fitness in steady state and hence is selected in nature in the very long run. To derive the conditions for maximum steady-state population growth, form the Lagrangian for optimization of $p(c, l, \beta)$ subject to the steady-state condition

$$(14) \quad c = f(M, k, h) - kp(c, l, \beta),$$

with total population and hence M taken as given. The first-order conditions for fitness-maximizing behavior simplify to

$$(15) \quad f_k = p(c, l, \beta),$$

$$(16) \quad \frac{1 - kp_c}{f_h + kp_l} = \frac{p_c}{p_l} = \frac{1}{f_h},$$

plus (14). The golden rule (15) is unchanged from the previous section. Condition (16) characterizes the fitness-maximizing mix of consumption and leisure. In addition to (14)–(16), the growth rate of total population equals the growth rate of the carrying capacity of the environment in very-long-run equilibrium:

$$(17) \quad p(c, l, \beta) = g.$$

For given β, the four conditions (14)–(17) determine the very-long-run equilibrium values k^*, l^*, c^*, and M^*.

Generation z is now represented as maximizing

$$(18) \quad U_z = \sum_{t=z}^{z+T} u(c_t, l_t, \beta) \times \prod_{v=z}^{t-1} \left(\frac{1 + p(c_v, l_v, \beta)}{1 + \delta} \right).$$

As before, the equilibrium solution is consistent with maximization of (18) if and only if

$$(19) \quad f_k = \delta.$$

Further, straightforward differentiation of (18) reveals that at c^* and l^*:

$$(20) \quad \frac{\partial U / \partial c_t}{\partial U / \partial l_t} = \frac{u_c + p_c \tilde{U}_t / (1 + \delta)}{u_l + p_l \tilde{U}_t / (1 + \delta)},$$

where

$$\tilde{U}_t \equiv \sum_{v=t+1}^{z+T} u(c_v, l_c, \beta) \left[(1 + p(c^*, l^*, \beta)) \Big/ (1 + \delta) \right]^{v-t-1}.$$

In the limit when $T \to \infty$, we have $\tilde{U}_t \to \infty$, so the marginal rate of substitution between consumption and leisure in naturally selected preferences equals the marginal rate of substitution in fitness, p_c / p_l. As before, the marginal rate of substitution between c_t and c_{t+1} in selected preferences equals the marginal rate of substitution between c_t and c_{t+1} as a determinant of population.

C. *Application to Development Economics*

To see how changes in the environment influence very-long-run behavior, we perform comparative statics on the equilibrium conditions (14)–(17). For brevity, we report the results given that $p(\cdot)$ is linear in c and l, or specifically that $p(\cdot)$ is of the form $\phi(\beta)c + \omega(\beta)l + \psi(\beta)$:

$$(21) \quad \frac{dM}{d\beta} = \frac{-p_\beta}{p_c f_M} < 0,$$

$$(22) \quad \frac{dh}{d\beta} = \left[\frac{p_{c\beta} f_h}{p_c} + \left(f_{Mh} + \frac{f_{kh} f_{Mk}}{-f_{kk}} \right) \frac{dM}{d\beta} \right] \Bigg/ D > 0,$$

$$(23) \quad \frac{dk}{d\beta} = \frac{f_{Mk}}{-f_{kk}} \frac{dM}{d\beta} + \frac{f_{kh}}{-f_{kk}} \frac{dh}{d\beta} > 0,$$

$$(24) \quad \frac{dc}{d\beta} = f_M \frac{dM}{d\beta} + f_h \frac{dh}{d\beta} > 0,$$

where $D = -[f_{kk}f_{hh} - (f_{kh})^2]/f_{kk} > 0$ because $f(\cdot)$ is concave. Equation (21) shows that an increase in environmental harshness as it is defined here reduces equilibrium population density. From equation (22), labor rises with environmental harshness because the relative value of consumption rises ($p_{c\beta} > 0$) and because the reduction in population density raises the marginal product of labor both directly (second term) and indirectly via increased capital (third term). Equation (23) verifies that capital also rises as a result of greater marginal productivity. Finally, equation (24) shows that per-capita consumption increases because population declines and per-capita labor rises. (The increase in per-capita capital has no impact on per-capita consumption at golden-rule saving.)

Remark 1: In broad terms, the method here allows classification of how the properties of the environment affect naturally selected preferences and thereby influence directly measurable magnitudes such as population density, capital, leisure, and material consumption. The calculations we report are an abbreviated illustration of the general method, however, as we impose specific restrictions on $p(\cdot)$ and $f(\cdot)$. That is, we assume that $p(\cdot)$ is linear in c and l, that per-capita output, $f(\cdot)$, does not depend directly on the environment (β), and that second-order cross derivatives among factors, including the effective amount of natural environment, are positive. Relaxing these assumptions would expand the number of effects encompassed by the analysis and make the signs of (21)–(24) ambiguous. If a harsher environment were to decrease output for given inputs, for instance, the tendencies in equations (21)–(24) might be strengthened. On the other hand, one can readily verify that if a harsher environment reduces the marginal product of capital greatly but changes the average product only slightly, capital may decline with environmental harshness. Thus generally, the validity of Toynbee's challenge and response hypothesis depends on how an increase in environmental challenge is assumed to influence fitness and production.

Remark 2: An interpretation of the analysis in this section is that selection in an agricultural environment with harsh winters may lead to selection of preferences that sustain an equilibrium with low population density and high per-capita capital, labor, and consumption. As suggestive examples, selected behavior (preferences) in such an environment might reflect strong ethics to work, to accumulate equipment and structures, to avoid depleting land, and to devote resources to upbringing and education of children.[19,20] If preferences that support high labor and capital are otherwise difficult to generate and if such preferences aid industrialization, then the analysis helps explain why industrialized countries are overrepresented in regions with harsh winters.[21] An addi-

[19] Cross-section evidence on OECD economies (Hansson and Stuart, 1988) suggests that the taste for labor is indeed greater in Northern Europe than in Southern Europe.

[20] In this spirit, one might expect greater educational expenditures in more northerly (harsher) areas. Rough data in UNESCO (1986, Table 4.1) indicate generally greater public expenditures on education as a percentage of GNP in Northern than Southern Europe. In 1975, for instance, these percentages were 7.8 in Denmark, 6.5 in Finland, 7.1 in Norway, and 7.1 in Sweden, but only 2.0 in Greece, 4.5 in Italy, 4.0 in Portugal, and 2.1 (in 1976) in Spain. (Because GNP per-capita is generally greater in Northern than in Southern Europe, the differences in educational expenditures as a share of GNP reported here understate differences in per-capita educational expenditures.) Data on direct general expenditures per capita by state and local governments in the United States for 1983–1984 (U.S. Department of Education, 1987, Table 25) provide a similar but perhaps weaker pattern. Counting as northern states those states that bound Canada or the Great Lakes (except for Indiana and Illinois, which bound southern Lake Michigan) and counting as southern states those states that bound Mexico or the Caribbean plus Hawaii, one calculates average per-capita educational expenditures across northern states of $901 and average expenditures across southern states of $728. One should probably interpret the latter numbers with caution, however, in part because it is unclear how long selection has been operating across the U.S. states.

[21] Two thousand years ago, the Mediterranean countries were the most developed in the Western hemisphere and Northern Europe was undeveloped. To make this consistent with the theory here, one would allow regional differences in production. The current treatment abstracts from such differences in that we assume that β does not directly influence per-capita output.

tional part of the picture, which is not incorporated into our model, is that greater per-capita capital (including human capital) may feed back positively on the rate of technological growth (Robert Lucas, 1985; Paul Romer, 1986). With such a feedback, environmental differences that lead to differences in tastes for work and saving might ultimately explain persistent regional differences in growth rates.

Remark 3: George Stigler and Becker (1977) argue that it is methodologically productive to view tastes as not differing "importantly between people" (p. 76). In our analysis, preferences selected in a given environment are homogeneous, so our very-long-run equilibrium has the property desired by Stigler and Becker. Preferences in small countries with little recent immigration may thus be quite homogeneous. Preferences in "melting-pot" countries such as the United States may be less homogeneous.

Remark 4: In this application, preferences for sacrificing own consumption to benefit offspring and preferences for the mix of own consumption and leisure are determined by natural selection. However, the idea that natural selection determines behavior represented by economists as the outcome of constrained optimization presumably has more general applicability. For instance, nothing restricts the theory here of sacrifice for progeny and work effort to *Homo sapiens*. To the extent that other species have shorter lives and ages of maturity, of course, selection of their economic behavior may be more rapid.

III. A General Statement

A general statement of the principle underlying this paper is that in very-long-run equilibrium, utility mirrors fitness in the sense that marginal rates of substitution in utility equal analogously defined marginal rates of substitution in fitness. To show this compactly, consider traits or preferences that determine the value of a vector x of behavioral variables that in turn determine fitness, $p(x)$. An element x_i of the vector can be interpreted as a commodity indexed by time and state of the world. (Note that x is of finite dimension here.) The individual or individual clan faces an individual resource constraint of the form $G(x,M)=0$, where M is total population deflated by the rate of growth of carrying capacity. An aggregate resource constraint $H(x,M)=0$ captures limits to the carrying capacity of the environment. In the absence of biological constraints on feasible traits, behavior in very-long-run equilibrium is characterized by maximization of $p(x)$ subject to $G(x,M)=0$ when M is held constant at a value that satisfies $H(x,M)=0$. Economists then represent behavior as the choice of x that maximizes $U(x)$ subject to $G(x,M)=0$. Assuming differentiability and comparing the two sets of first-order conditions, it must be that $U_{x_i}/U_{x_j}=p_{x_i}/p_{x_j}$, or that in very-long-run equilibrium, marginal rates of substitution in the naturally selected behavioral utility function equal marginal rates of substitution in fitness written as a function of behavioral variables. Note that this result does not rely on an assumption that agents actually maximize utility. One may conveniently *describe* naturally selected behavior in very-long-run equilibrium as the outcome of utility maximization even if individuals do not actually maximize utility.

IV. Conclusion

We use natural selection to explain behavior, or equivalently, preferences. We focus specifically on intergenerational saving and on consumption-leisure choice, using a neoclassical golden-age model in which population is endogenized through classical Malthusian assumptions about the economic determinants of population size. The model depicts a world in which different "clans" have different preferences and hence different relative rates of population growth or fitness. In the very-long-run, the economy rests in an equilibrium in which (i) 100 percent of the population have preferences with maximum fitness given resource constraints and (ii) total population equals the carrying capacity of the environment. A generic characteristic of equilibrium is that marginal rates of substitution in utility equal marginal rates of substitution in fitness written as a function of behavioral variables. We show that selection implies a tendency toward golden-rule saving, which means that agents act as if they maximize the undiscounted sum of per-capita felicities of current and future generations. Selection also predicts that population density, work, saving, and consumption should vary across regions in response to regional differences in the natural environment.

REFERENCES

Abel, Andrew and Kotlikoff, Laurence, "Does the Consumption of Different Age Groups Move Together? A New Nonparametric Test of Intergenerational Altruism," Boston University Working Paper No. 164, 1988.

Alchian, Armen A., "Uncertainty, Evolution, and Economic Theory," *Journal of Political Economy*, June 1950, *58*, 211–21.

Barro, Robert J., "Are Government Bonds Net Wealth?" *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

Becker, Gary S., "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology," *Journal of Economic Literature*, September 1976, *14*, 817–26.

______, *A Treatise on the Family*, Cambridge, MA: Harvard University Press, 1981.

______ **and Lewis, H. Gregg,** "On the Interaction Between the Quantity and Quality of Children," *Journal of Political Economy*, March 1973, *81*, S279–88.

Boyd, Robert and Richerson, Peter, *Culture and the Evolutionary Process*, Chicago: University of Chicago Press, 1985.

Boyer, George, "Malthus Was Right After All: Poor Relief and Birth Rates in Southeastern England," *Journal of Political Economy*, February 1989, *97*, 93–114.

Coale, Ansley, "The History of the Human Population," *Scientific American*, September 1974, *231*, 40–51.

Frank, Robert, "If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review*, September 1987, *77*, 593–604.

Hamilton, W. D., "The Genetical Evolution of Social Behavior," *Journal of Theoretical Biology*, July 1964, *7*, 1–52.

Hansson, Ingemar and Stuart, Charles, "Taxes and Aggregate Labor Supply: A Cross-Country General-Equilibrium Study," U.C. Santa Barbara Working Paper No. 10, 1988.

______ **and** ______, "Social Security as Trade Among Living Generations," *American Economic Review*, December 1989, *79*, 1182–95.

Hirshleifer, Jack, "Economics from a Biological Perspective," *Journal of Law and Economics*, April 1977, *20*, 1–52.

Huizinga, John and Mishkin, Frederic S., "The Measurement of Short-Term Real Interest Rates on Assets with Different Risk Characteristics," Working Paper, 1983.

Knodel, John, *The Decline of Fertility in Germany, 1871–1939*, Princeton, NJ: Princeton University Press, 1974.

Kotlikoff, Laurence J. and Summers, Lawrence H., "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 706–32.

Lee, Ronald, "Population Dynamics of Humans and Other Animals," *Demography*, November 1987, *24*, 443–65.

Leibowitz, Arleen, "Home Investments in Children," *Journal of Political Economy*, March 1974, *82*, S111–31.

Lucas, Robert E., "On the Mechanics of Economic Development," reproduced, University of Chicago, 1985.

Lumsden, Charles and Wilson, Edward O., *Genes, Mind, and Culture*, Cambridge, MA: Harvard University Press, 1981.

Malthus, T. R., *Essays on Population*, 1798.

Margolis, Howard, *Selfishness, Altruism, and Rationality*, Cambridge: Cambridge University Press, 1982.

Maynard Smith, John, *Evolution and the Theory of Games*, Cambridge: Cambridge University Press, 1982.

Mincer, Jacob, "Market Prices, Opportunity Costs, and Income Effects," in Carl Christ et al., eds., *Measurement in Economics*, Stanford, CA: Stanford University Press, 1963.

______, *Schooling, Experience, and Earnings*, New York: National Bureau of Economic Research, 1974.

Modigliani, Franco, "Life Cycle, Individual Thrift, and the Wealth of Nations," *American Economic Review*, June 1986, *76*, 297–313.

______ **and Brumberg, Richard,** "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in K. Kurihara, ed., *Post-Keynesian Economics*, New Brunswick, NJ: Rutgers University Press, 1954.

Romer, Paul, "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*, 1002–38.

Rubin, Paul and Paul, Chris, "An Evolutionary Model of Taste for Risk," *Economic Inquiry*, October 1979, *17*, 585–96.

Stigler, George J. and Becker, Gary S., "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, *67*, 76–90.

Toynbee, Arnold J., *A Study of History*, New York: Oxford University Press, 1956.

Willis, Robert J., "A New Approach to the Economic Theory of Fertility Behavior," *Journal of Political Economy*, March 1973, *81*, S14–64.

Wilson, Edward O., *Sociobiology*, Cambridge, MA: Harvard University Press, 1980.

Wrigley, E. A. and Schofield, R. S., *The Population History of England, 1541–1871*, London: Edward Arnold, 1981.

U.S. Department of Education, *Digest of Education Statistics, 1987*, Washington: USGPO, 1987.

UNESCO, *Statistical Yearbook, 1986*, Gembloux, Belgium, 1986.

Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States

By LESLIE A. WHITTINGTON, JAMES ALM, AND H. ELIZABETH PETERS*

The effect of the income tax system has been largely ignored in existing empirical work on fertility.[1] This neglect is surprising. It has often been demonstrated that tax impacts on behavior can be quite substantial, even in areas in which economic costs and benefits are not commonly thought to play a dominant role.[2] Moreover, the governments of many countries act as if they believe that they can affect the birthrate through tax incentives. A concern about population decline has induced some governments to adopt explicitly pronatalist policies (Michael S. Teitelbaum and Jay M. Winter, 1985). France and West Germany have extensive systems of family allowances. East Germany and Hungary have implemented policies that include one-time birth payments and paid maternity leave (Henk J. Heeren, 1982, Jerome S. Legge, Jr., and John R. Alford, 1986). Canada and Singapore also attempt to influence fertility rates through tax policies (Malcolm Gillis, Dwight H. Perkins, Michael Roemer, and Donald R. Snodgrass, 1983). At the other extreme, China has received international attention for the extreme antinatalist economic incentives that it has instituted. A few studies report that these policies have some impact on fertility rates, but the magnitude of that impact is still an issue (Kingley Davis, Mikhail S. Bernstam, and Rita Ricardo-Campbell, 1986).

The United States has not implemented such explicit policies as family allowances or paid maternity leaves. However, the federal income tax has a feature that may implicitly affect the decision to have a child: the personal exemption for dependents. As noted by Joseph Pechman (1983), the personal exemption is justified not as a policy for influencing the fertility decision, but as relief for low-income households and families of the burden of taxation; its amount is roughly based on the income needed to maintain an adequate diet. Nevertheless, the personal exemption also represents a clear subsidy for each child, a subsidy whose value depends upon the marginal tax rate of the family. The effect of this tax feature on the aggregate fertility rate in the United States has not been previously examined.

In this paper, we estimate an aggregate fertility equation for the United States from 1913 to 1984. Fertility is modeled as a function of various economic and demographic factors, including the tax value of the personal exemption. The primary result is that the personal exemption has a positive and significant effect on the national birthrate, and this result is robust to a variety of specifications. Although the elasticity of the birthrate with respect to the exemption is not large, it appears that the United States —and perhaps other countries as well—can influence to a degree the fertility decisions of its citizens through deliberate changes in tax policies.

*Assistant Professor, Department of Textiles and Consumer Economics, University of Maryland, College Park, MD 20742, Assistant Professor, Department of Economics, University of Colorado, Boulder, CO 80309-0256, and Assistant Professor, Department of Economics and Research Associate, Population Program, University of Colorado, Boulder, CO 80309-0256. The authors would like to thank Steve Elliott, Robert Hutchens, Robert McNown, the participants of seminars at the University of Colorado, North Carolina State University, University of Louisville, Trinity College, and the University of Michigan-Dearborn, and three referees for helpful comments.

[1] A few economists have examined this issue. Peter H. Lindert (1978) estimates child costs over time and includes the tax benefits as an input to those costs. T. Paul Schultz (1981) speculates that the tax system might affect the demand for children. More recently Thomas Espenshade and Joseph J. Minarik (1987) estimate the impact on fertility of the 1986 reform act. In their analysis, however, they ignore the potential effect of a change in the cost of a child and focus, instead, on income effects.

[2] This literature is enormous. A useful starting point is Henry J. Aaron and Joseph Pechman (1981).

TABLE 1—CHILD COSTS AND THE PERSONAL EXEMPTION FOR DEPENDENTS

Study	Annual Child Costs Medium Income Group[a]	Tax Value of Personal Exemption[b]	Personal Exemption As a Percent of Annual Expense in Percent
Turchi (1983)			
(1981 Dollars)			
Avg Male Child	$3470.76	$325.CO	9.36
Espenshade (1984)			
(1981 Dollars)			
One Child	5900.04	325.00	5.51
2nd Child Addition	3255.48	325.00	9.98
3rd Child Addition	2311.20	325.00	14.06
Olson (1983)			
(1982 Dollars)			
One Male Child	7293.12	295.00	4.04
2nd Child Addition	5015.16	295.00	5.88
BLS (1982)			
(1981 Dollars)			
One Child	3405.00	325.00	9.54
2nd Child Addition	2938.44	325.00	11.06
3rd Child Addition	2958.84	325.00	10.98
USDA (1982)			
(1981 Dollars)			
Avg Urban Child	4466.64	325.00	7.28

[a] Child costs derived from Williams (1987), II-145. (Monthly expenses × 12 months)
[b] In 1981 the value of the personal exemption is 1,000 dollars per dependent and the average marginal tax rate is 32.5 percent; in 1982 the personal exemption is 1,000 dollars per dependent and the average marginal tax rate is 29.5 percent.

Section I examines the impact of the personal exemption on the costs of a child. Section II presents the empirical model of fertility, and discusses the data and variables that are used. Section III presents the empirical results. Conclusions and policy implications are summarized in Section IV.

I. Child Costs and the Personal Exemption

The standard economic model of fertility posits that children provide their parents with positive utility in either consumption or production. The cost of a child depends upon the cost of time inputs and the cost of the goods and service inputs used in child rearing.[3] The cost of a child also depends upon the annual tax savings generated by that child, equal to the value of the personal exemption multiplied by the marginal tax rate of the parent claiming the child as a dependent. This annual subsidy rises with the marginal tax rate and with the size of the personal exemption.

The value of the personal exemption is not inconsequential relative to the annual cost of raising a child. Robert G. Williams (1987) has compiled data on child costs from a number of economic studies. As shown in Table 1, these estimates vary considerably. However, it is obvious that in all cases the tax subsidy represents a large portion of the annual monetary expenses of child rearing. Using a medium socioeconomic status as a measure, the tax value of the personal exemption represents from 4 to 9 percent of the annual estimated monetary cost of raising one child. For subsequent children, the tax value of the personal exemption increases as a portion of the annual child costs, ranging from almost 6 to 14 percent.

[3] See Gary Becker (1960) and Schultz (1973) for a theoretical discussion. For examples of empirical studies linking fertility and economic variables, see William P. Butz and Michael P. Ward (1979), Marc Nerlove and T. Paul Schultz (1970), Michael Wachter (1975).

It is also important to recognize that the personal exemption is an ongoing support item. The parents of a child receive this tax subsidy for every year that the child is filed as a dependent. In most cases this will be a minimum of eighteen years. Thus the personal exemption is a stream of subsidies to birth, not a one-shot payment. At a discount rate of 10 percent, the present value of an annual stream of payments of $325—the average tax value of the personal exemption in 1981—is nearly $3000; at a discount rate of 5 percent the present value rises to almost $4000.

The personal exemption assumes a much different role than other economic subsidies to children that have been examined. For example, the relationship between birthrates and various income maintenance programs like Aid to Families with Dependent Children (AFDC) has been explored, and the results generally fail to prove that higher welfare payments lead to higher fertility.[4] AFDC, however, is a system of payments extended only to those with economic need, usually only to single parent families, and is not necessarily received over the entire dependency period of the child.

It is clear that the personal exemption encourages fertility by decreasing the relative cost of children. Of course, whether and how much individuals in fact respond to this incentive is an empirical issue.

II. The Empirical Framework

The previous section argued that the cost of children affects the fertility decision, and that these costs depend in part upon the tax value of the personal exemption. However, there are clearly other factors that influence the birth decision. This section discusses those other factors and presents the basic model of fertility.

We estimate the following reduced form equation for the period 1913, the year in which the modern federal income tax came into existence, to 1984:

$$\begin{aligned}(1)\quad \text{General Fertility Rate} &= \beta_0 + \beta_1 \text{Personal Exemption} \\ &+ \beta_2 \text{Income} \\ &+ \beta_3 \text{Unemployment} \\ &+ \beta_4 \text{Infant Mortality} \\ &+ \beta_5 \text{Immigration} + \beta_6 \text{Female Wage} \\ &+ \beta_7 \text{Birth Control} + \beta_8 \text{World War II} \\ &+ \beta_9 \text{Time Trend}.\end{aligned}$$

Table 2 gives the variables, definitions, and means.

The dependent variable, the general fertility rate, is the birthrate per thousand women between the ages of 15 and 44. This is the group commonly considered to be at risk of pregnancy. The general fertility rate is less sensitive to changes in the age and sex structure of the population than the crude birthrate. The fertility rate series is reported in Appendix 1.

Due to biological constraints, the birth of a child will lag the decision to have a child. For this reason we estimate the fertility equation in lagged form. Several different lag structures are used to test the sensitivity of our results to assumptions about the timing of the fertility decision process. The correct lag structure is difficult to identify. One appealing structure is an inverted V pattern with weights initially increasing and then decreasing. The rationale for this form comes from the biological average of 24 to 31 months required to produce a birth (T. Paul Schultz, 1981). Under this structure, the variable is lagged four periods from the current period, and the peak comes in the $t-2$ period. The constructed variable W_{it} becomes:

$$(2)\qquad W_{it} = w_1 X_t + w_2 X_{t-1} + w_3 X_{t-2} + w_4 X_{t-3} + w_5 X_{t-4},$$

[4] See Glen G. Cain (1977) and David T. Ellwood and Mary Jo Bane (1985).

TABLE 2—VARIABLE DEFINITIONS AND MEANS

Variable Name	Definition	Mean	Standard Deviation
General Fertility Rate	Births per 1000 women aged 15 to 44	95.50	19.64
Personal Exemption	Real tax value of personal exemption	100.40	65.88
Male & Asset Income	Real after-tax personal income per family net of female earnings	7466.37	2982.78
Unemployment	Unemployment rate of civilian labor force	0.071	0.053
Infant Mortality	Infant mortality per 1000 live births	43.02	26.84
Immigration	Immigration of at-risk group as a percent of resident at-risk group	0.003	0.0035
Female Wage	Average after-tax female wage	1.22	0.532
Pill	Dummy variable equal to one in years 1963 to 1984	0.305	0.464
WW II	Dummy variable equal to one in years U.S. was in World War II	0.069	0.256
Time Trend	Time trend equal to one in 1913 and increasing by one each year	36.50	20.92
Female Education	Female high school graduates each year as a percent of female population	0.009	0.004

where $w_1 < w_2 < w_3$ and $w_3 > w_4 > w_5$. The lag structure thus declines around the mean lag of two years, which is the average time required to produce a birth. A two-year, rectangular lag structure is also examined:

$$(3) \qquad W_{it} = w_1X_t + w_2X_{t-1} + w_3X_{t-2},$$

where $w_1 = w_2 = w_3$. Several other lag structures have been estimated with no significant impact on the results.

The independent variable of primary interest is the real tax value of the personal exemption, equal to the personal exemption multiplied by the average marginal tax rate. The federal income tax began in 1913, and the personal exemption became a feature of the federal tax system in 1917. We deflate values of the personal exemption using the Consumer Price Index. We use average marginal tax rates estimated by Robert J. Barro and Chaipat Sahasakul (1983, 1986).[5] Although Congress has changed the personal exemption only nine times between 1913 and 1984, the real tax value of the exemption exhibits substantial fluctuation due to changes in the average marginal tax rate and in the price level. See Figure 1 and Appendix 1 for this series.

Additional independent variables include both those that affect birthrates by changing

[5] Barro and Sahasakul (1983, 1986) calculate the average marginal tax rate from 1916 to 1983. Their methodology is employed to calculate the rate for 1984. Though the empirical work reported here looks at values 1913 to 1984, there was no personal exemption prior to 1917. Thus no values of PE are missing. Barro and Sahasakul report average marginal tax rate series weighted by adjusted gross income and by number of returns filed; both approaches are calculated arithmetically and geometrically. All four series were tried in estimating the model with no substantial difference in the results.

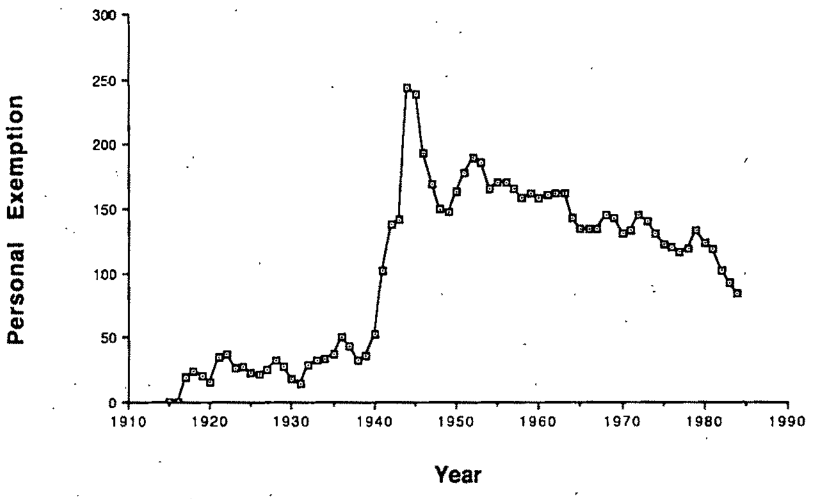


FIGURE 1. REAL TAX VALUE OF PERSONAL EXEMPTION

the demand for children and biological factors that influence the supply of births. Income is measured as the sum of male earnings and non-wage family property income.[6] Because women in general still commit more time to the care of children than do men (Arleen Leibowitz, 1975), women's wages are a reasonable proxy for the time cost of children. Consequently, if mother's earnings are included in family income, the estimated coefficient on that variable confounds the estimates of the income effect with those of a price effect. To estimate a pure income effect, we therefore consider family income net of mother's earnings. If children are a normal good, we expect the sign on income to be positive. However, Gary Becker and H. Gregg Lewis (1973) argue that high income families may invest more in the quality of each child, so that income may have a negative effect on quantity.

Measuring the cost of time inputs presents major difficulties. In fact, the lack of an adequate measure of the value of women's time may be partially responsible for the relatively small number of time-series studies on fertility. Women's wages were not regularly reported in the first half of the century, so there is no continuous wage series available for statistical work. In addition, as much of the cross-section work on fertility during the last decade has noted, the observed market wage is a biased measure of the value of time for women who do not work in the market.[7] This paper constructs a wage series from various sources. June O'Neill (1985) reports the ratio of female to male wages from 1955 to 1982. Using these ratios and data on production wages, we construct a female wage series. We then multiply wages by the average marginal tax rate to obtain the after-tax female wage. See Appendix 2 for details about the construction of this series. Though the series represents only one sector, wage trends across sectors over time are highly correlated.

The infant mortality rate is included to capture two possible effects. The death of a child could cause an increase in the birthrate if families are concerned about completed family size; this is commonly called the replacement effect. However, infant mortality increases the cost of producing a surviving child. If the cost effect dominates, then an increasing infant mortality rate could lower the birthrate.

[6] Male earnings are from the Statistical Abstract, various years, and Current Population Reports, Series P-60. Average property income is derived from the national income accounts.

[7] See, for example, James Heckman (1974).

Unemployment may also have an ambiguous effect on fertility rates. Unemployment results in lower transitory income, so that high rates of unemployed workers would lower birthrates if children are a normal good. However, unemployment is also likely to lower the opportunity cost of spending time in the production of children. The time effect would result in a higher birthrate. Because of its transitory nature, we predict that unemployment will primarily affect the timing of births rather than the number of ever-born children.

Birthrates also differ across cultures. Changing immigration rates in the United States may therefore account for a portion of the change in U.S. fertility over time.

A dummy variable equal to one during World War II (1941–45) accounts for the absence of young men during the war years, and a dummy variable equal to one for the years that the birth control pill has been widely available (1963–1984) is also included. Finally, we use several specifications of a time trend to capture any unobserved socioeconomic factors that might affect fertility and that have changed over time.

III. Estimation Results

Table 3 reports estimation results. Generalized least squares estimation is performed with a Yule-Walker first-order autocorrelation correction scheme. Six models are presented, including those with differing lag structures. Figure 2 compares the actual fertility rate with the rate predicted from model 4. In general, the fit is good. However, actual rates are slightly higher than predicted during the peak baby boom years and somewhat lower than predicted during the years of the Great Depression.

Of primary concern here is the impact of the real tax value of the personal exemption. The personal exemption has a positive and significant impact on birthrates in all specifications in Table 3. The results are robust to the specifications of different lag structures, models 1–4, to the exclusion of a time trend, model 5, and to the substitution of female education for female wages, model 6. The coefficients range in value from 0.121 to 0.236, and the estimated elasticities of fertility with respect to the personal exemption range from 0.127 to 0.248. These estimated elasticities are consistent with the wage elasticities of fertility estimated in our study and smaller than those reported in other time-series studies.[8] The largest values occur with a five-period inverted V lag structure. These results suggest that an increase in the tax value of the personal exemption of 50 dollars will increase the general fertility rate by 6 to 12 births per 1000 women at risk.

Perhaps surprisingly, the coefficient on real personal income per household is often negative, although not significantly different from zero. The negative sign on income is a fairly common result in cross-sectional tests.[9] With a few exceptions such as Marc Nerlove and T. Paul Schultz (1970), time-series studies have generally found a positive impact.[10]

Other variables generally have predicted signs. The coefficients on the infant mortality rate are positive, indicating that in the United States the replacement effect dominates the cost effect, but the coefficient is rarely statistically significant. The positive coefficient is consistent with the estimates by Schultz (1974) in Taiwan, Nerlove and Schultz (1970) in Puerto Rico, and Michael P. Shields and Ronald L. Tracy (1986) in the United States. Unemployment has a negative effect on childbirth, which suggests that the temporary income effect dominates the value of time effect of unemployment.

The performance of the measure of female wages is generally disappointing; the sign is usually negative, but the coefficient is often insignificantly different from zero. It might be argued that female wages and fertility are simultaneously determined, so that inclusion of a female wage creates endogeneity problems. We tested for endogeneity using the Jerry A. Hausman (1978) specification test. This test consisted of regressing the sus-

[8] Our wage elasticities range from 0.03 to 0.18. Butz and Ward (1979) report a wage elasticity of 0.751 to 1.846.

[9] See also Julian Simon, 1969.

[10] See, for example, Wachter (1975), Butz and Ward (1979), Shields and Tracy (1986).

TABLE 3—IMPACT OF THE PERSONAL EXEMPTION ON FERTILITY IN THE UNITED STATES, 1913–1984

	Model					
Independent Variable	1	2	3	4	5	6
Personal Exemption	0.121**	0.191**	0.230**	0.157**	0.236**	0.211**
	(2.714)	(4.040)	(4.805)	(3.959)	(5.018)	(4.56)
Male & Asset Income	−0.0004	0.0004	−0.0004	−0.0019	−0.0005	0.001
	(0.147)	(0.141)	(0.174)	(0.760)	(0.258)	(0.644)
Unemployment	−73.430**	−36.862	−31.455	−30.328	−34.639	−16.828
	(2.147)	(1.108)	(0.998)	(1.000)	(1.112)	(0.526)
Infant Mortality	0.083	0.303	0.310	0.296	0.439**	0.180
	(0.325)	(1.263)	(1.284)	(1.300)	(2.470)	(0.750)
Immigration	774.24**	1529.20**	1372.87**	319.55	1433.43**	1181.48**
	(2.487)	(3.183)	(3.119)	(1.244)	(3.324)	(2.723)
Female Wage	5.647	−2.157	−8.712	−11.261	−13.804	–
	(0.360)	(0.153)	(0.661)	(0.867)	(1.200)	
Pill	−10.856*	−8.958	−6.922	−5.561	−7.854	−5.952
	(1.772)	(1.622)	(1.364)	(1.067)	(1.599)	(1.197)
WW II	−17.223**	−10.449**	−5.353	0.016	−4.997	−4.771
	(3.452)	(2.596)	(1.356)	(0.004)	(1.279)	(1.240)
Time Trend	−0.539	−0.389	−0.377	0.025	–	−0.471
	(1.002)	(0.785)	(0.803)	(0.051)		(1.171)
Female Education	–	–	–	–	–	−2196.46*
						(1.781)
Intercept	102.979**	79.961**	81.628**	93.978**	74.488**	93.270**
	(4.175)	(3.373)	(3.531)	(4.114)	(3.519)	(3.975)
R^2	0.916	0.931	0.941	0.943	0.940	0.944
Elasticity of Fertility with Respect to Exemption	0.127	0.201	0.242	0.165	0.248	0.221

Absolute value of *t*-statistic in parentheses.

Model 1: No lags on independent variables.

Model 2: Three-year rectangular lag on personal exemption; no lags on other independent variables.

Model 3: Five-year inverted *V* on personal exemption; no lags on other independent variables.

Model 4: Two-year lag on all independent variables except pill, World War II and time trend.

Model 5: Five-year inverted *V* on personal exemption; no lags on other independent variables; no time trend.

Model 6: Five-year inverted *V* on personal exemption; no lags on other independent variables, female education used as a proxy for female wage.

**: Significant at the 5 percent level.
*: Significant at the 10 percent level.

pected endogenous variable-female wages-on all exogenous variables in the fertility regression plus the variables education, race, and urban population. The actual value of the wage and the residual from this regression were then included in a fertility regression. Endogeneity is not a significant problem if the *t*-statistic on the residual is insignificant. Identification was achieved by including variables—education, race, and urban population—in the wage equation that are excluded from the fertility equation. The *t*-statistic on the wage residual was 0.555, which indicates that endogeneity is not a severe problem. As an additional test of the sensitivity of our results to the possibility of endogenous wages, Table 3 reports a reduced form fertility regression that uses female education as a proxy for female wages (model 6).

We examined a variety of other specifications, and the results are generally consistent with those presented in Table 3.[11] The results are robust to the inclusion of additional variables such as the racial distribution of

[11] These results are available upon request.

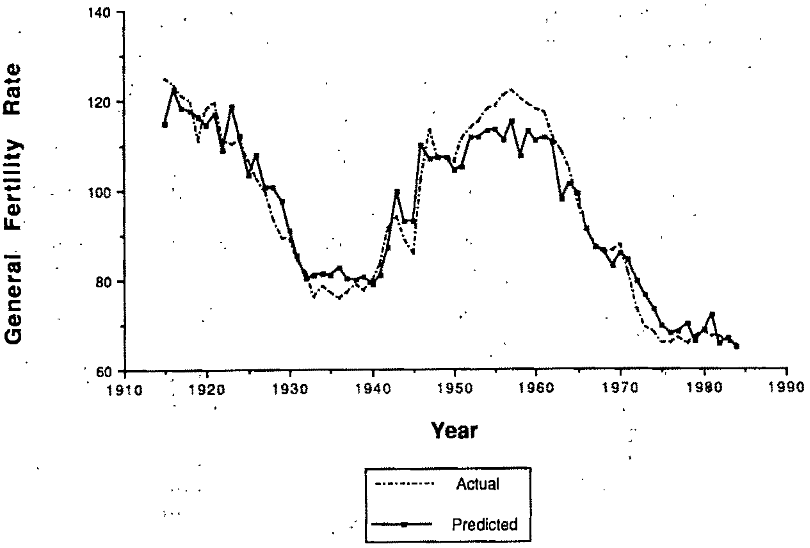


FIGURE 2. GENERAL FERTILITY RATE—ACTUAL AND PREDICTED

the population and the percent of urban population. These variables, however, did not contribute significantly to the explanatory power of the equations, and they were dropped from the analysis. The results are also robust to various lag structures for the independent variables and to the specification of various time trends, (for example, no time trend, a linear time trend, and a quadratic time trend). We also tested for structural differences in the estimates across different subperiods. Regressions done separately for periods 1913–48 and 1949–84 both yielded positive and significant coefficients on the tax value of the personal exemption. A Chow test on this split sample indicated that there was no structural difference in the full regression and the sub-regressions.[12] Because the largest variations in the tax value of the personal exemption occurred during a short period of time, 1940–49, the subperiod results are somewhat sensitive to the particular years included in each subperiod. On the whole, however, these results support the hypothesis that an increase in the tax value of the personal exemption leads to an increase in the demand for children.

IV. Implications for Public Policy

Our empirical results indicate that tax policy, at least in the form of the personal exemption, has an impact on aggregate family birth decisions. The policy ramifications for this particular tax feature are potentially important. Consider, for example, what our estimates imply about the impact of the Tax Reform Act (TRA) of 1986 on the birthrate. The TRA changes the tax value of the personal exemption in several ways. The statutory value of the personal exemption increases from $1,080 in 1986 to $2,000 in 1989 when fully phased in; thus the nominal value of the personal exemption virtually doubles. However, this increase is dampened because marginal tax rates are reduced for many taxpayers.[13] The overall impact will depend on whether the increasing nominal value of the personal exemption or the falling marginal tax rate dominates. Further, the personal exemption is fully applicable only to taxpayers with taxable incomes of amounts less than $149,250; the birth subsidy is not extended to the highest income households. Finally, a large number of

[12] The Chow test yielded on F-statistic of 0.420 with 8 and 56 degrees of freedom, which is less that the critical value of 1.69.

[13] Jerry A. Hausman and James Poterba (1987) calculate that 47.7 percent of taxpayers will have a marginal tax rate decrease of 0 to 10 percent, and 11.3 percent will have a decrease greater than 10 percent.

households will no longer be within the tax system under the new tax law. The zero bracket amount is increased substantially; thus the lowest income groups will be outside of the tax system.

Consider a married couple filing jointly with income of $16,000. Prior to TRA, this couple received a personal exemption per dependent of $1,080 and faced a marginal tax rate of 16 percent; the tax value of the personal exemption in 1986 dollars was therefore $173. In 1989, earning the same income, this couple now would receive a personal exemption per dependent of $2,000 and face a marginal tax rate of 15 percent. Assuming a 4 percent rate of inflation, the tax value of their personal exemption in 1986 dollars would now be $267, an increase of almost 55 percent. A family who faced a marginal tax rate of 49 percent on taxable income of $109,400 prior to TRA would experience a much smaller change in the tax value of the personal exemption. Formerly, this subsidy had a value to them of $529. Under TRA this subsidy is worth $533, an increase of less than 1 percent. The average increase in the tax value of the personal exemption is fifteen percent.

Overall, the estimates in this paper indicate that the TRA changes in the personal exemption may generate a significant increase in the birthrate. Using the most conservative coefficient estimate of the impact of the personal exemption on birthrates (0.121), the increased personal exemption when fully phased in will cause an increase of 7.53 births per thousand women at risk. This represents an 11 percent rise in the current birthrate. Middle income families will receive the largest birth incentive, while low and high income groups will experience a disincentive to births. Although the elasticity of births with respect to the personal exemption is small, statutory changes in the value of this subsidy have usually been relatively large. In particular, the changes due to TRA are longer than the changes during all but a few other years in our time-series. Thus our estimated fertility response is substantial.

Of course, there are other aspects of TRA that might mitigate the fertility effect of an increase in the personal exemption. For example, the lower marginal tax rates instituted by the law will increase the opportunity cost of spending time with children and may lower the demand for children. In addition, if TRA leads to an increase in after-tax income, and if the income elasticity for child quality is large, the demand for numbers of children may decline. In this paper we do not address these other factors and thus the total impact of TRA on fertility is likely to be smaller than the estimates that we present here. Because of the potentially large impact of this tax policy tool on fertility rates, empirical work using micro structures data or data from countries with different tax structures would be useful to provide additional evidence about the size of this impact.

APPENDIX 1

General Fertility Rate and the Personal Exemption for Dependents, 1913–1984

Year	Birthrate	Real Tax Value of PE
1913	124.7	0
1914	126.6	0
1915	125.0	0
1916	123.4	0
1917	121.0	19.27
1918	119.8	23.94
1919	111.2	20.07
1920	117.9	15.33
1921	119.8	34.32
1922	111.2	36.65
1923	110.5	25.83
1924	110.9	27.34
1925	106.6	22.85
1926	102.6	21.13
1927	99.8	24.61
1928	93.8	31.96
1929	89.2	27.29
1930	89.2	18.40
1931	84.6	14.91
1932	81.7	28.36
1933	76.3	31.95
1934	78.5	33.91
1935	77.2	36.98
1936	75.8	50.12
1937	77.1	42.79
1938	79.1	32.22
1939	77.6	36.53
1940	79.9	53.33
1941	83.4	102.49
1942	91.5	137.70
1943	94.3	141.20
1944	88.4	243.83
1945	85.9	238.40
1946	101.9	193.16

APPENDIX 1—Continued

Year	Birthrate	Real Tax Value of PE
1947	113.3	168.90
1948	107.3	149.79
1949	107.1	147.05
1950	106.2	163.10
1951	111.5	178.14
1952	113.9	189.43
1953	115.2	186.51
1954	118.1	165.46
1955	118.5	170.57
1956	121.2	171.00
1957	122.9	165.12
1958	120.2	158.66
1959	118.8	162.19
1960	118.0	158.28
1961	117.2	160.71
1962	112.2	161.58
1963	108.5	161.61
1964	105.0	142.73
1965	96.6	134.60
1966	91.3	133.94
1967	87.6	133.80
1968	85.7	145.10
1969	86.5	142.62
1970	87.9	130.58
1971	81.8	132.99
1972	73.4	144.85
1973	69.2	140.87
1974	68.4	130.49
1975	66.0	122.36
1976	65.8	120.08
1977	66.8	116.11
1978	65.5	118.98
1979	67.2	132.93
1980	68.4	123.17
1981	67.4	119.31
1982	67.3	102.04
1983	65.8	92.49
1984	65.4	83.90

APPENDIX 2

Average Female Real Wage, 1913–1984
(1967 = 1.00)

Year	Wage	Year	Wage
1913	0.461	1924	0.738
1914	0.458	1925	0.712
1915	0.467	1926	0.713
1916	0.492	1927	0.717
1917	0.503	1928	0.747
1918	0.554	1929	0.737
1919	0.547	1930	0.738
1920	0.627	1931	0.735
1921	0.657	1932	0.702
1922	0.681	1933	0.786
1923	0.720	1934	0.972

APPENDIX 2–Continued

Year	Wage	Year	Wage
1935	0.959	1960	1.776
1936	0.928	1961	1.739
1937	0.981	1962	1.777
1938	0.988	1963	1.812
1939	1.000	1964	1.855
1940	1.043	1965	1.903
1941	1.084	1966	1.859
1942	1.147	1967	1.918
1943	1.278	1968	1.979
1944	1.351	1969	2.063
1945	1.358	1970	2.064
1946	1.359	1971	2.057
1947	1.368	1972	1.094
1948	1.405	1973	2.061
1949	1.323	1974	2.034
1950	1.239	1975	2.103
1951	1.235	1976	2.170
1952	1.287	1977	2.187
1953	1.423	1978	2.277
1954	1.404	1979	2.206
1955	1.661	1980	2.136
1956	1.669	1981	2.106
1957	1.729	1982	2.173
1958	1.746	1983	2.216
1959	1.765	1984	2.240

Construction of Wage Series Years. 1914, 1920–1948: From Historical Statistics, Series D830–844, Average Hourly Earnings of Production Workers by Sex.

Years 1913, 1915–1919: Estimated based on average wage growth in production, from Series 802–810 and Series 765–778.

Years 1955–1982: O'Neill (1985) calculates a female/male wage ratio based on annual wage earnings and average hours worked by both sexes. This ratio is transformed into a female/average wage ratio, so that information on average production wages can be used to determine female production wages. Using the relationship:

$$\text{Female/Male} = (\text{Average/Male}) \times (\text{Female/Average})$$

the female/average ratio can be calculated from O'Neill's (1985) existing series on female/male wages and data on male wages and average wages found in Statistical Abstract, various years. This ratio is adjusted upward by 0.027, to account for the difference between O'Neill's ratio and the observed female/male wage ratio in production in 1939, a year in which both are available. Finally, the adjusted female/average ratio is multiplied by the average wage in manufacturing to arrive at a female wage rate.

Years 1949–1954, 1983–1984: Using median wage earnings, and adjusting for an approximate 9 percent hour differential between men and women, the ratio of female to male wages is calculated. This is transformed into a female/average ratio using the techniques described above. This ratio is multiplied by the average manufacturing wage.

We compared this series to a decennial series provided by Claudia Goldin. This series trends in the same way as Goldin's series.

REFERENCES

Aaron, Henry J. and Pechman, Joseph A., eds., *How Taxes Affect Economic Behavior*, Washington: The Brookings Institution, 1981.

Barro, Robert J. and Sahasakul, Chaipat, "Measuring the Average Marginal Tax Rate from the Individual Income Tax," *Journal of Business*, October 1983, *56*, 419–52.

______, "Average Marginal Tax Rates from Social Security and the Individual Income Tax," *Journal of Business*, October 1986, *59*, 555–66.

Becker, Gary, "An Economic Analysis of Fertility," in *Demographic and Economic Change in Developed Countries*, Princeton, NJ: Princeton University Press for the NBER, 1960.

______ **and Lewis, H. Gregg,** "On the Interaction Between the Quantity and Quality of Children," *Journal of Political Economy*, March/April 1973, *81*, S279–S288.

Butz, William P. and Ward, Michael P., "The Emergence of Countercyclical U.S. Fertility," *American Economic Review* June 1979, *69*, 318–28.

Cain, Glen G., "Fertility Behavior," in *The New Jersey Income Maintenance Experiment*, Vol. III, Harold W. Watts and Albert Rees, eds., New York: Academic Press, 1977, 225–50.

Calhoun, Charles A. and Espenshade, Thomas J., "Childbearing and Wives' Foregone Earnings," *Population Studies*, *42*, March 1988, 5–37.

Davis, Kingley, Bernstam, Mikhail S. Ricardo-Campbell, Rita, eds., *Below-Replacement Fertility in Industrial Societies*, Cambridge: Cambridge University Press, 1986.

Ellwood, David T. and Bane, Mary Jo, "The Impact of AFDC on Family Structure and Living Arrangements," in *Research in Labor Economics*, Vol. 7, Ronald G. Ehrenberg, ed. Greenwich: JAI Press, 1985, 137–207.

Espenshade, Thomas and Minarik, Joseph J., "Demographic Implications of the 1986 U.S. Tax Reform," *Population and Development Review* March 1987, *13*, 115–27.

Gillis, Malcolm, Perkins, Dwight H., Roemer, Michael and Snodgrass, Donald R., *Economics of Development*, New York: W. W. Norton & Co., 1983.

Gronau, Reuben, "The Intrafamily Allocation of Time: The Value of the Housewife's Time," *American Economic Review*, September 1973, *63*, 634–51.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–71.

______ **and Poterba, James,** "Household Behavior and the Tax Reform Act of 1986," *Journal of Economic Perspectives*, Summer 1987, *1*, 110–19.

Heckman, James, "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, July 1974, *42*, 679–94.

Heeren, Henk J., "Pronatalist Population Policies in Some Western European Countries," *Population Research and Policy Review*, 1982, *1*, 137–52.

Legge, Jr., Jerome S. and Alford, John R., "Can Government Regulate Fertility? An Assessment of Pronatalist Policy in Eastern Europe," *Western Political Quarterly*, December 1986, *39*, 709–28.

Leibowitz, Arleen, "Education and the Allocation of Women's Time," in *Education, Income and Human Behavior*, F. Thomas Juster, ed., New York: McGraw-Hill, 1975, 171–97.

Lindert, Peter H., *Fertility and Scarcity in America*, Princeton, NJ: Princeton University Press, 1978.

Nerlove, Marc and T. Paul, Schultz, *Love and Life Between the Census: A Model of Family Decision Making in Puerto Rico, 1950–1960*. Santa Monica: Rand Corporation, RM-6322-AID. September 1970.

O'Neill, June, "The Trend in the Male-Female Wage Gap in the United States," *Journal of Labor Economics*, January 1985 Supplement, *3*, S91–S116.

Pechman, Joseph, *Federal Tax Policy*, Washington: The Brookings Institution, 1983.

Schultz, T. Paul, "A Preliminary Survey of Economic Analyses of Fertility," *American Economic Review*, May 1973, *63*, 71–87.

______, "Birth Rate Changes over Space and Time: A Study of Taiwan," in *Economics of the Family*, Theodore W. Schultz, ed. Chicago: University of Chicago Press, 1974, 255–91.

______, *Economics of Population*, Reading, MA: Addison Wesley, 1981.

Shields, Michael P. and Tracy, Ronald L., "Four Themes in Fertility Research," *Southern Economic Journal*, July 1986, *53*, 201–16.

Simon, Julian, "The Effect of Income on Fertility," *Population Studies*, November 1969, *23*, 327–41.

Teitelbaum, Michael S. and Winter, Jay M., *The Fear of Population Decline*, Orlando, FL: Academic Press, 1985.

Wachter, Michael, "A Time Series Fertility Equation: The Potential for a Baby-Boom in the 1980's," *International Economic Review*, October 1975, *61*, 609–24.

Williams, Robert G., *Development of Guidelines for Child Support Orders: Final Report*, U.S. Department of Health and Human Services, Office of Child Support Enforcement, March 1987.

The Relationship Between the Marginal Cost of Public Funds and Marginal Excess Burden

By Robert K. Triest*

A large literature has recently developed which attempts to measure the marginal cost (shadow price) of public funds.[1] The motivation for much of this literature is the recognition that many of the tax instruments used by governments generate substantial amounts of excess burden (deadweight loss). Jerry A. Hausman (1981), for example, estimates that for an "average" prime-aged American male with a wage of ten dollars in 1975 the excess burden generated by the federal income tax was over fifty percent of the revenue raised from taxing that individual. One might suspect that such large excess burden estimates would require a change in the way government project costs are calculated in benefit-cost analysis. If the revenue required to finance a project is raised through distortionary taxation, then intuition would suggest that the excess burden generated by the increase in tax rates needed to finance the project should be incorporated into the cost of the project.

However, this analysis is not entirely consistent with the theoretical literature on the optimal provision of public goods which are financed by distortionary taxation. When distortionary taxes are used to finance public goods, and the level of public good provision does not affect consumption of the taxed goods, the standard Samuelsonian rule which equates the sum (over consumers) of the marginal rates of substitution between a private good and a public good with the marginal rate of transformation between the two goods must be modified to account for the marginal effect of increases in tax rates on the revenue raised from the existing taxes.[2] The intuition behind this result is that we need to correct for indirect "leakages" from existing government revenue sources when considering an incremental expenditure. If the increase in tax rates needed to finance the project causes a reduction in the revenue raised by the current tax structure, then we need to account for this loss in calculating the cost of the project.

It is important to note that this result concerns the values of uncompensated demand elasticities. An example first provided by Atkinson and Stern (1974) illustrates this point. Consider a world where the only tax is on labor earnings, and incremental expenditure on the public good does not affect the demand for leisure. In this case, the marginal rate of transformation will overstate the cost of the public good as long as the uncompensated wage elasticity of labor supply is negative. This is because an increase in the tax rate causes an increase in labor supply, leading to an indirect increase in government revenue.

From the excess burden perspective described earlier, this example is very puzzling. It is well known that even if the uncompensated own price elasticity of a good is zero, a tax on that good can still result in a large deadweight loss. We instead need the compensated elasticity to be zero to ensure that there will be no excess burden. However, the

*Department of Economics, University of California, Davis, CA 95616. This paper is based on chapter two of my dissertation at the University of Wisconsin-Madison. I wish to thank the members of my dissertation committee, Martin David, Robert Haveman (chair), and Arthur Goldberger, for very helpful advice and comments. I have also benefited from comments by Don Fullerton and Bruce Hamilton. An anonymous referee provided unusually helpful comments and corrected an error in the figure.

[1] Don Fullerton (1989) provides a recent survey and analysis of this literature.

[2] This is shown by Anthony Atkinson and Nicholas H. Stern (1974) and David E. Wildasin (1979). The effect of provision of the public good on consumption of the taxed goods must also generally be taken into account.

labor supply example seems to indicate that the cost of an incremental project might need to be adjusted downward to account for the effects of distortionary taxation even when the marginal excess burden generated by the taxes is positive.

The purpose of this paper is to provide an explanation of this puzzle. The basic argument is that the most appropriate measure of excess burden is based on a price vector which differs from the one which would be most commonly used to measure benefits in benefit-cost analysis. Since excess burden and benefit-cost analysis are based on different price vectors, an index number correction must be made in order to relate the two concepts. Due to this, the marginal cost of public funds may be less than one even under a tax system with large associated values of both total and marginal excess burden.

Section I of the paper briefly describes a measure of excess burden and evaluates how it relates to benefit-cost analysis. In Section II, the relationship between marginal excess burden and the marginal cost of public funds is derived. Section III provides a numerical example which indicates that marginal excess burden may be a very misleading indicator of the marginal cost of public funds. Section IV concludes the paper with a discussion of the results.

I. Excess Burden and Benefit-Cost Analysis

Although the basic idea behind excess burden is very simple, there is a large literature on the appropriate way to measure it. A recent review of this literature is provided by Alan J. Auerbach (1985). In order to simplify the analysis, we will consider only the case of a single consumer economy with fixed producer prices. The fixed producer prices assumption is equivalent to assuming that the production possibilities frontier is linear. These assumptions allow us to ignore problems of income distribution and producer price changes caused by government policy.

Measures of excess burden attempt to measure, in monetary units, the decrease in the consumer's utility due to imposition of a tax system minus the value of the real resources which are captured by the government through the tax system. In other words, we want to measure to what degree the loss of the consumer's welfare due to the tax system exceeds the loss which would have resulted from the use of a non-distortionary (lump-sum) tax. In the case of a lump-sum tax, the utility loss and the value of the resources captured by the government are equal and offsetting. However, when distortionary taxes are used, the decrease in consumer welfare exceeds the revenue raised by the tax system.

The various measures of excess burden which have been proposed differ in (a) how the change in the consumer's utility is measured and (b) how the magnitude of the tax revenue is measured. A concensus appears to have formed that the most acceptable way of measuring excess burden is to subtract the actual revenue raised by the tax system from the equivalent variation associated with the imposition of the tax system. This measure, which was proposed independently by John A. Kay (1980) and Elisha A. Pazner and Efraim Sadka (1980), may be written algebraically as

$$(1)\qquad e(\mathbf{q}, u_1) - e(\mathbf{p}, u_1) - \mathbf{t}\cdot\mathbf{x}(\mathbf{q}, y),$$

where $\mathbf{p}$ is the pretax (producer) price vector, $\mathbf{t}$ is the vector of commodity taxes, $\mathbf{q}$ is the post-tax price vector (which is equal to $\mathbf{p}+t$), $e(\mathbf{p}, u)$ is the expenditure function evaluated at price vector $\mathbf{p}$ and utility level u, and $x(\mathbf{p}, y)$ is the vector of Marshallian demands evaluated at prices $\mathbf{p}$ and income level y. The utility level at which the expenditure function is evaluated, u_1, is the level of well-being corresponding to the post-tax world; it may be written in terms of the indirect utility function as $u_1 = v(\mathbf{q}, y)$. Verbally, equation (1) may be interpreted as the difference between the maximum amount the consumer would have been willing to pay to avoid the imposition of the tax system and the amount of revenue raised. This measure has the desirable properties of decreasing due to any tax reform which either (i) holds tax revenue constant, but results in a higher level of post-tax utility, or (ii) keeps post-tax utility constant, but results in an increase in tax revenue. In addition, it will be zero if only lump-sum taxes are imposed and will be positive for any tax system which is less efficient than lump-sum taxation.

One reason why the Kay-Pazner-Sadka excess burden measure has such desirable properties is that both the welfare loss and tax revenue are measured in terms of producer prices. The equivalent variation (EV) can be rewritten as the difference between the pretax and post-tax utility levels measured in terms of producer prices:

$$(2)\qquad \begin{aligned} \mathrm{EV} &= e(\mathbf{q}, u_1) - e(\mathbf{p}, u_1) \\ &= y - e(\mathbf{p}, u_1) \\ &= e(\mathbf{p}, u_0) - e(\mathbf{p}, u_1). \end{aligned}$$

Similarly, tax revenue is equal to the value, in terms of producers prices, of the difference between the bundle of goods consumed without taxes and the bundle of goods consumed after the tax system is imposed:

$$(3)\qquad \begin{aligned} \mathbf{t}\cdot\mathbf{x}(\mathbf{q}, y) &= (\mathbf{q}-\mathbf{p})\cdot\mathbf{x}(\mathbf{q}, y) \\ &= \mathbf{q}\cdot\mathbf{x}(\mathbf{q}, y) - \mathbf{p}\cdot\mathbf{x}(\mathbf{q}, y) \\ &= \mathbf{p}\cdot[\mathbf{x}(\mathbf{p}, y) - \mathbf{x}(\mathbf{q}, y)] \end{aligned}$$

(where the last step follows from $\mathbf{p}\cdot\mathbf{x}(\mathbf{p}, y) = \mathbf{q}\cdot\mathbf{x}(\mathbf{q}, y) = y$). It would be inappropriate to substitute compensating variation (CV) for equivalent variation in the excess burden formula, since compensating variation is a measure of the value of the change in utility in terms of the final post-tax price vector:

$$(4)\qquad \begin{aligned} \mathrm{CV} &= e(\mathbf{q}, u_0) - e(\mathbf{p}, u_0) \\ &= e(\mathbf{q}, u_0) - e(\mathbf{q}, u_1). \end{aligned}$$

As pointed out by Kay (1980), Pazner and Sadka (1980), and Wilfried Pauwels (1986), difficulties arise when excess burden is defined using the compensating variation to measure the decrease in welfare due to taxation. The reason for this is that while compensating variation is based on the post-tax price vector, tax revenue is a measure of the value of resources based on the pretax price vector. Because of this, combining the two into a sensible measure of excess burden is very difficult.

Recently, several authors have generalized the Hicksian consumer's surplus measures to the case where a monetary measure of the distance between u_0 and u_1 is obtained in terms of any fixed price vector:

$$(5)\qquad \begin{aligned} &e(\mathbf{r}, u_0) - e(\mathbf{r}, u_1) \\ &\quad = e(\mathbf{r}, v(\mathbf{p}, y)) - e(\mathbf{r}, v(\mathbf{q}, y)), \end{aligned}$$

where $\mathbf{r}$ is any constant reference price vector.[3] Picking $\mathbf{r}$ equal to $\mathbf{p}$ (the original price vector) yields the equivalent variation. Setting $\mathbf{r}$ to $\mathbf{q}$ (the final consumer price vector) yields the compensating variation. As long as $\mathbf{r}$ is held constant over all alternative final price vectors being compared, this expression is a consistent ordinal indicator of the utility associated with these alternatives (this follows from $e(\mathbf{r}, u_0)$ being constant over all alternatives and $e(\mathbf{r}, u_1)$ being increasing in u_1).

In performing a benefit-cost analysis of a proposed project, an explicit choice must be made for the reference price vector. In actual benefit-cost studies, net benefits are normally based on the current distorted consumer price vector rather than on the producer price vector. For example, an individual's answer to a question asking how much she would be willing to pay to have a particular project built is given in terms of the prices that the consumer currently faces rather than the prices she would face if all distortions were removed. While any price

[3] Angus Deaton (1980) and Hal R. Varian (1984) present this measure in their discussion of consumer's surplus and show that the compensating and equivalent variations are special cases of it. Mervyn A. King (1983) calls this measure the "welfare gain."

vector can serve as the reference vector, measures based on current consumer prices have the important advantage of being easily interpretable by both the policy analysts and those affected by the project. If net benefits are measured in terms of current consumer prices, then in order to be consistent the measure of the project's cost should also be in terms of these prices.

Suppose that the current tax vector is $\mathbf{t}_1$, the consumer price vector is $\mathbf{q}_1$ ($\mathbf{q}_1 = \mathbf{p}+\mathbf{t}_1$), and a change in the tax vector from $\mathbf{t}_1$ to $\mathbf{t}_2$ is introduced in order to finance a new project ($\mathbf{q}_2 = \mathbf{p}+\mathbf{t}_2$). A appropriate measure of the cost of the project in terms of current prices is

$$(6)\quad e(\mathbf{q}_1, v(\mathbf{q}_1, y_i)) - e(\mathbf{q}_1, v(\mathbf{q}_2, y_i)).$$

This is the equivalent variation (starting from price vector $\mathbf{q}_1$) associated with the imposition of the additional taxes needed to pay for the project.[4] The cost measure (equation (6)) is closely related to the first two terms of the Kay-Pazner-Sadka measure of excess burden (equation (1)). If we change the reference price vector of the cost measure from $\mathbf{q}_1$ to $\mathbf{p}$, then it becomes identical to the first two terms of the excess burden measure. The key difference is that excess burden is based on producer prices, while the cost measure for use in benefit-cost analysis is based on current consumer prices.

Figure 1 illustrates the importance of the choice of the reference price vector.[5] Suppose a consumer allocates his income between two goods, x_1 and x_2. The pretax budget constraint is AE and the consumer is initially at point A'. After a tax is imposed on x_1, the budget constraint becomes AF and the consumer moves to point B'. The revenue raised by this tax (letting x_2 serve as numeraire) is AB. The equivalent variation associated with the imposition of the tax is AC. Therefore, the deadweight loss of the tax (using the Kay-Pazner-Sadka measure) is BC. Notice that even though consumption of the taxed good does not change due to the imposition of the tax, a positive quantity of deadweight loss is generated.

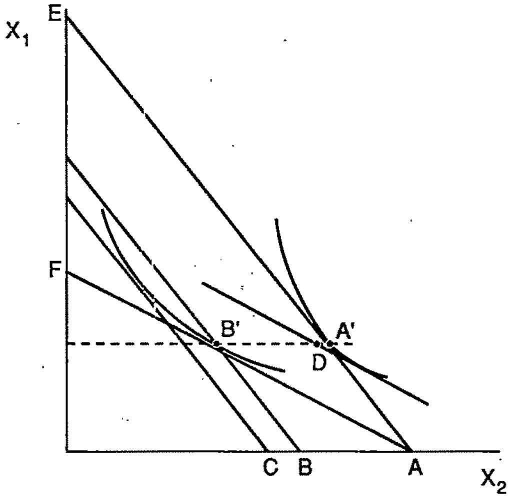


FIGURE 1

The loss in consumer welfare due to the imposition of the tax in terms of post-tax prices is DB'. This measure of welfare loss is less than the amount of revenue raised ($A'B'$). However, as noted earlier, this does not imply that there is no deadweight loss since the welfare loss and tax revenue are now based on different sets of reference prices.

In calculating the marginal cost of public funds to use in benefit-cost analysis where benefits are calculated in terms of current consumer prices, we need to find how much the welfare cost measure based on the distorted price vector (DB') increases when tax revenue is increased by one dollar. In the next section, it is demonstrated that in this example the marginal cost of public funds is equal to one. When the taxed good has an uncompensated own price elasticity of zero, the marginal increase in the welfare cost measure is exactly equal to the extra revenue raised.

[4] See Triest (1987b) for analogous benefit and cost measures which explicitly model the introduction of the new public good.

[5] This figure is similar to Figure 1 in Pazner and Sadka (1980), where they demonstrate that basing excess burden on compensating variation may lead to incorrect welfare rankings.

II. Marginal Excess Burden and the Marginal Cost of Public Funds

The marginal excess burden associated with a particular tax is usually defined as the increase in excess burden which results from a one-dollar increase in tax revenue generated through use of that tax. Suppose the current tax vector is $\mathbf{t}_2$ (with consumer price vector $\mathbf{q}_2$). The marginal excess burden associated with increasing the tax on good i is then the partial derivative of the excess burden measure with respect to the tax on good i divided by the partial derivative of tax revenue with respect to the tax on good i:

$$(7)\quad \frac{\left[\dfrac{\partial\left[\mathbf{e}(\mathbf{p},\mathbf{v}(\mathbf{p},y)) - e(\mathbf{p},\mathbf{v}(\mathbf{q}_2,y)) - \mathbf{t}_2\cdot\mathbf{x}(\mathbf{q}_2,y)\right]}{\partial q_{2i}}\right]}{\left[\dfrac{\partial[\mathbf{t}_2\cdot\mathbf{x}(\mathbf{q}_2,y)]}{\partial t_{2i}}\right]} = \frac{\left[\dfrac{-\partial e(\mathbf{p},\mathbf{v}(\mathbf{q}_2,y))}{\partial q_{2i}}\right]}{\left[\dfrac{\partial[\mathbf{t}_2\cdot\mathbf{x}(\mathbf{q}_2,y)]}{\partial t_{2i}}\right]} - 1.$$

This expression is closely related to the marginal cost of public funds. The marginal cost of public funds may be defined as the increase in the cost measure which results when an additional dollar of tax revenue is raised. If the benefit-cost analyst treats required tax revenue as the cost of the project, then the marginal cost of public funds is the minimum required benefit-cost ratio for incremental expenditures. As we will see below, the marginal cost of public funds depends critically on the price vector which benefits are measured in terms of.

When benefits are measured in terms of some reference price vector $\mathbf{r}$ and revenue is raised through an increase in the tax on good i, the appropriate measure of the marginal cost of public funds is the partial derivative of equation (6) with respect to t_i divided by the partial derivative of tax revenue with respect to t_i (this equation is derived in the Appendix):

$$(8)\quad -\frac{\left[\dfrac{\partial \mathbf{e}(\mathbf{r}, v(\mathbf{q}_2, y))}{\partial \mathbf{q}_{2i}}\right]}{\left[\dfrac{\partial[\mathbf{t}_2\cdot\mathbf{x}(\mathbf{q}_2,y)]}{\partial t_{2i}}\right]} = \frac{\left[\dfrac{\partial v(\mathbf{q}_2,y)}{\partial y}\right]}{\left[\dfrac{\partial v(\mathbf{r},y-\mathrm{EV})}{\partial y}\right]} \left[1+\frac{\sum_j t_{2j}\dfrac{\partial x_j(\mathbf{q}_2,y)}{\partial q_{2i}}}{x_i(\mathbf{q}_2,y)}\right]^{-1},$$

where EV is defined implicitly by $v(\mathbf{r}, y-\mathrm{EV}) = v(\mathbf{q}_2, y)$. Equation (8) is an appropriate expression for the marginal cost of public funds for a project whose benefits will be measured in terms of price vector $\mathbf{r}$. When producer prices ($\mathbf{p}$) are used as the reference prices in the project evaluation, then the marginal cost of public funds is equal to one plus marginal excess burden.

However, as argued in the preceding section, benefit-cost analysis is rarely done in terms of undistorted prices. When benefits and costs are instead measured in terms of the final price vector, $\mathbf{q}_2$, the first term in the above expression becomes one. In this situation, the marginal cost of public funds depends only on the ratio of the indirect loss of tax revenue due to the marginal increase in the price of good i,

$$\sum_j t_{2j}\frac{\partial x_j(\mathbf{q}_2,y)}{\partial q_{2i}},$$

to the direct increase in tax revenue, $x_i(q_2, y)$. This measure appears throughout the literature, although its relationship to marginal excess burden and the choice of $\mathbf{q}_2$ as the reference price vector has not been noted.[6] When benefits are measured in terms of prices which differ from $\mathbf{q}_2$, the first term acts as an index number adjustment for the differences in the marginal utility of income under the two price vectors.

In Section I, it was argued that the prices facing consumers at the time the project evaluation is performed, $\mathbf{q}_1$, is normally used as the reference price vector in benefit-cost analysis. For a small project, the pre-project and post-project consumer price vectors will be approximately the same ($\mathbf{q}_1 \approx \mathbf{q}_2$). In this case, the marginal cost of public funds will be greater than one only to the extent that the increase in the tax rates used to generate revenue results in a decrease in the weighted (by tax rates) sum of the uncompensated demands for the taxed goods.

When marginal excess burden is calculated on the basis of the Kay-Pazner-Sadka measure, $\mathbf{p}$ is used as the reference price vector rather than $\mathbf{q}_2$. In this case, the marginal cost of public funds corresponding to the $\mathbf{q}_2$ reference price vector must be multiplied by the ratio of

$$\frac{\partial v(\mathbf{q}_2, y)}{\partial y} \quad \text{over} \quad \frac{\partial v(\mathbf{p}, y-\mathrm{EV})}{\partial y}.^{7}$$

The inverse of the marginal utility of income serves the purpose of quantifying the level of utility associated with a given price vector. The second term in the marginal cost of public funds expression is a measure of the value, in terms of current consumer prices ($\mathbf{q}_2$), of the indirect loss in welfare due to raising an additional dollar of revenue through increasing the tax on good i. In order to convert this to a measure of welfare loss in terms of price vector $\mathbf{p}$, we must multiply by the ratio of the actual marginal utility of income over what the marginal utility of income would be if the consumer faced price vector $\mathbf{p}$. The equivalent variation is subtracted from the income term in $v(p, y-\mathrm{EV})$ so that the marginal utility of income is evaluated at the actual current level of utility in both the numerator and denominator of the ratio.

The marginal cost of public funds in terms of reference price vector $\mathbf{p}$ will exceed one if

$$\frac{\partial v(\mathbf{q}_2, y)}{\partial y}$$

is greater than

$$\frac{\partial v(\mathbf{p}, y-\mathrm{EV})}{\partial y}$$

even if the increase in the tax in question causes no reduction in the demand for any of the taxed goods.[8] An intuitive explanation for this phenomenon is that the dollar value of a given reduction in welfare will be greater in terms of $\mathbf{p}$ than in terms of $\mathbf{q}_2$ whenever each dollar of income is worth more to the consumer under prices $\mathbf{q}_2$ than under prices $\mathbf{p}$. Therefore, if the marginal utility of income is greater in terms of price vector $\mathbf{q}_2$ than it is in terms of price vector $\mathbf{p}$, then we must adjust upward a measure of marginal welfare change measured in terms of prices $\mathbf{q}_2$ in order to convert it to a measure of value in terms of price vector $\mathbf{p}$.

III. A Numerical Example

A numerical example suggests that changing the reference price vector can make a great deal of difference. Suppose there are two goods, consumption (the numeraire) and leisure (with relative price w). The con-

[6] This is the same as the expression derived by Atkinson and Stern (1974) and Wildasin (1979) when the increase in the level of public good provision accompanying the tax increase is ignored. Neville Topham (1984, 1985) attempts to link the marginal cost of public funds to marginal excess burden, but unfortunately bases his analysis on excess burden measures based on either the compensating variation or Marshallian consumer's surplus. Similarly, Edgar K. Browning (1987) bases his measure of the marginal cost of public funds on the compensating variation.

[7] Where EV is defined implicitly by $v(\mathbf{q}_2, y) = v(\mathbf{p}, y-\mathrm{EV})$.

[8] A similar point is made by Pazner and Sadka (1980).

TABLE 1—MARGINAL COST OF PUBLIC FUNDS SIMULATIONS PARAMETER SETTINGS[a]

	$\alpha = 11.3$ $\beta = 0.113$ $\gamma = 2663$	$\alpha = 11.3$ $\beta = 0.565$ $\gamma = 3115$	$\alpha = 56.5$ $\beta = 0.113$ $\gamma = 2527$	$\alpha = 56.5$ $\beta = 0.565$ $\gamma = 3318$
Wage Elasticities[b]				
Uncompensated	0.01	0.01	0.07	−0.07
Compensated	0.35	1.71	0.40	1.63
Marginal Cost of Public Funds:				
Reference Wage = \$5 (Gross Pretax Wage)	1.26	3.12	1.31	2.97
Reference Wage = \$4	1.13	1.77	1.17	1.69
Reference Wage = \$3 (Net Post-tax Wage)	1.01	1.10	1.05	0.96
Excess Burden (Kay-Pazner-Sadka Measure):	665.0	4450.0	753.0	4245.0

[a] The parameters were chosen so that in all simulations \$5167.80 revenue is raised. See the text for further details.

[b] All elasticities were evaluated at $w = \$3.00$ and $Y = \$1000$.

sumer's preferences are represented by the indirect utility function:

$$(9)\quad v(w,Y) = e^{\beta w}\left(Y + \frac{\alpha}{\beta}w - \frac{\alpha}{\beta^2} + \frac{\gamma}{\beta}\right)$$

where Y is unearned income and α, β, and γ are parameters. The implied expenditure and labor supply functions are

$$(10)\quad e(w,u) = e^{-\beta w}u - \frac{\alpha}{\beta}w + \frac{\alpha}{\beta^2} - \frac{\gamma}{\beta},$$

and

$$(11)\quad h(w,Y) = \gamma + \alpha w + \beta Y.$$

This is the specification used by Hausman (1981). For an "average" prime-aged married male, Hausman estimated γ to be approximately 2663, α to be 11.3, and β to be -0.113 (where h is annual hours of work, and w and Y are both measured in dollars).[9] All calculations using this model were performed assuming a gross wage of \$5 and gross unearned income of \$1000.

Table 1 shows the marginal cost of public funds for a slight increase in the income tax starting from a tax rate of forty percent (net wage of \$3). The calculations were done for four different combinations of parameters in order to see how the results vary with the labor supply elasticities. The parameters were picked such that the same amount of revenue is raised in each simulation. Three different reference prices (wages) were used: the gross wage, the final net wage, and an intermediate value. Since the numerator of the marginal cost of public funds expression (equation (8)) is equal to one when the net wage (\$3) is used as reference wage, the numerator for any other reference wage can be easily calculated by taking the ratio of the marginal cost of public funds with that reference wage to the marginal cost of public funds with the reference wage set at \$3.

As one would expect, the excess burden of the tax is positive for all four sets of parameter values. Excess burden is largest in the simulations with high compensated wage elasticities. Marginal excess burden may be quickly calculated from this table by subtracting one from the marginal cost of public funds figures which use the pretax wage as the reference wage. Again as expected, the two simulations with relatively large compensated wage elasticities also have the largest values of marginal excess burden.

[9] The γ estimate is conditional on certain demographic characteristics; see Gary Burtless' comment on Hausman (1981, p. 76) for details. The β estimate is actually the median of an estimated distribution for β.

The marginal cost of public funds calculated using the net post-tax wage as the reference wage follows a pattern quite different from marginal excess burden. For the parameter values in the first column, which are based on Hausman's (1981) labor supply estimates, the marginal cost of public funds is 1.26 when it is based on marginal excess burden (using the pretax wage as the reference wage), but only 1.01 when using the net post-tax wage.

The most striking example in this table is the results of the simulation reported in column four. In this case, the compensated wage elasticity is fairly large, while the uncompensated elasticity is negative. When the undistorted wage is used as the reference price, the marginal cost of public funds is approximately three. However, when the net wage is used as the reference price, the marginal cost drops below one. Thus, there is the potential for large errors in using a marginal cost of public funds calculated in terms of prices which differ from those which are used to measure benefits. Even if there is a large marginal excess burden (in terms of the Kay-Pazner-Sadka excess burden measure) associated with increasing the tax on a good, the marginal cost of public funds (when benefits are measured in terms of post-tax prices) associated with raising revenue from that tax may be less than one.

As the numerical example shows, the marginal cost of public funds may be very sensitive to the choice of the reference price vector. In doing applied welfare analysis, it is very important to use a single set of reference prices for all quantities being compared. Using a marginal excess burden measure based on one reference price vector to calculate the marginal cost of public funds when benefits are calculated using a different reference price vector may produce very misleading results.

IV. Concluding Remarks

This paper has demonstrated the importance of carefully linking the marginal cost of public funds to the benefits measure actually used in project evaluation. A given estimate of the marginal cost of public funds is of no use unless it is shown how it relates to a particular framework for project evaluation. Estimates of the marginal cost of public funds based on unusual welfare measures must be treated cautiously since it is hard to relate them to standard benefit-cost analysis.

When benefits are measured in terms of current consumer prices, then the marginal cost of public funds depends only on uncompensated price elasticities. This result is surprising, since excess burden depends on compensated demand curves. This paper shows that the reason for this paradoxical result is that excess burden is normally defined in terms of undistorted prices. An index number problem then arises when trying to use marginal excess burden to calculate the marginal cost of public funds, which must be based on current consumer prices.

As noted by previous authors, even if the only taxed good has a negative uncompensated own-price elasticity, the marginal cost of public expenditure may exceed one if the additional expenditure results in a decrease in the demand for the taxed good. However, this is quite different from the usual excess burden argument. When public expenditures increase consumption of taxed goods at the margin, then the public expenditure effect will work toward driving the marginal cost of public expenditure below one.

APPENDIX

Derivation of Equation (8):

In order to derive equation (8), we must differentiate an expenditure function with respect to one of the price arguments of the indirect utility function it is defined over. Triest (1987a) shows that the following relationship holds:

$$\frac{\partial e(\mathbf{p}, v(\mathbf{q}, y))}{\partial q_i}$$

$$= -x_i(\mathbf{q}, y)\frac{\left[\dfrac{\partial v(\mathbf{q}, y)}{\partial y}\right]}{\left[\dfrac{\partial v(\mathbf{p}, y-\mathrm{EV})}{\partial y}\right]},$$

where EV is defined implicitly by $v(\mathbf{q}, y) = v(\mathbf{p}, y-\mathrm{EV})$.

A proof of this relationship is included here for the sake of completeness:

The first step is to show that $\partial e(\mathbf{p}, v(\mathbf{q}, y))/\partial v = 1/[\partial v(\mathbf{p}, y-\mathrm{EV})/\partial y]$:

$$e(\mathbf{p}, v(\mathbf{q}, y)) = e(\mathbf{p}, \mathrm{v}(\mathbf{p}, y-\mathrm{EV}))$$

$$= y - \mathrm{EV}$$

$$\frac{\partial e(\mathbf{p},\mathrm{v}(\mathbf{p},\mathrm{y}-\mathrm{EV}))}{\partial y}$$

$$= \frac{\partial e(\mathbf{p},\mathrm{y}(\mathbf{p},\mathrm{y}-\mathrm{EV}))}{\partial v}\,\frac{\partial v(\mathbf{p}, y-\mathrm{EV})}{\partial y} = 1$$

$$\frac{\partial e(\mathbf{p}, v(\mathbf{q}, y))}{\partial v} = \frac{\partial e(\mathbf{p}, v(\mathbf{p}, y-\mathrm{EV}))}{\partial v}$$

$$\frac{\partial e(\mathbf{p}, v(\mathbf{q}, y))}{\partial v} = \frac{1}{\left[\dfrac{\partial v(\mathbf{p}, y-\mathrm{EV})}{\partial y}\right]}.$$

We can now use this result to derive $\partial e(\mathbf{p}, v(\mathbf{q}, y))/\partial q_i$:

$$\frac{\partial e(\mathbf{p}, v(\mathbf{q}, y))}{\partial q_i} = \frac{\partial e(\mathbf{p}, v(\mathbf{q}, y))}{\partial v}\,\frac{\partial v(\mathbf{q}, y)}{\partial q_i}$$

$$= \frac{\left[\dfrac{\partial v(\mathbf{q}, y)}{\partial q_i}\right]}{\left[\dfrac{\partial v(\mathbf{p}, y-\mathrm{EV})}{\partial y}\right]}$$

multiplying and dividing by $[\partial v(\mathbf{q}, y)/\partial y]/[\partial v(\mathbf{p}, y-\mathrm{EV})/\partial y]$ yields (by Roy's identity):

$$= -x_i(\mathbf{q}, y)\frac{\left[\dfrac{\partial v(\mathbf{q}, y)}{\partial y}\right]}{\left[\dfrac{\partial v(\mathbf{p}, y-\mathrm{EV})}{\partial y}\right]}. \qquad \square$$

We can now use this identity to derive equation (8):

$$-\frac{\left[\dfrac{\partial e(\mathbf{r}, v(\mathbf{q}_2, y))}{\partial q_{2i}}\right]}{\left[\dfrac{\partial [\mathbf{t}_2\cdot\mathbf{x}(\mathbf{q}_2, y)]}{\partial t_{2i}}\right]}$$

$$= \frac{\left[x_i(\mathbf{q}_2, y)\dfrac{\left[\dfrac{\partial v(\mathbf{q}_2, y)}{\partial y}\right]}{\left[\dfrac{\partial v(\mathbf{r}, y-\mathrm{EV})}{\partial y}\right]}\right]}{\left[\sum_j t_{2j}\dfrac{\partial x_j(\mathbf{q}_2, y)}{\partial q_{2i}} + x_i(\mathbf{q}_2, y)\right]}$$

(where EV is implicitly defined by $v(\mathbf{q}_2, y) = v(\mathbf{r}, y + \mathrm{EV})$)

$$= \frac{\left[\dfrac{\partial v(\mathbf{q}_2, y)}{\partial y}\right]}{\left[\dfrac{\partial v(\mathbf{r}, y-\mathrm{EV})}{\partial y}\right]}\left[1 + \frac{\sum_j t_{2j}\dfrac{\partial x_j(\mathbf{q}_2, y)}{\partial q_{2i}}}{x_i(\mathbf{q}_2, y)}\right]^{-1}.$$

REFERENCES

Atkinson, Anthony and Stern, Nicholas H., "Pigou, Taxation and Public Goods," *Review of Economic Studies*, January 1974, *41*, 119–28.

Auerbach, Alan J., "The Theory of Excess Burden and Optimal Taxation," in Alan J. Auerbach and Martin Feldstein, eds., *The Handbook of Public Economics*, Vol. 1, New York: North-Holland, 1985.

Ballard, Charles L., Shoven, John, B. and Whalley, John, "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States," *American Economic Review*, March 1985, *75*, 128–38.

Browning, Edgar K., "On the Marginal Welfare Cost of Taxation," *American Economic Review*, March 1987, *77*, 11–23.

Burtless, Gary, "Comments" on paper by Jerry A. Hausman, in Henry J. Aaron and Joseph A. Pechman, eds., *How Taxes Affect Economic Behavior*, Washington: The Brookings Institution, 1981.

Deaton, Angus, "The Measurement of Welfare: Theory and Practical Guidelines," LSMS Working Paper No. 7, The World Bank, 1980.

Fullerton, Don, "If Labor Is Inelastic, Are Taxes Still Distorting?" National Bureau of Economic Research Working Paper No. 2810, January 1989.

Hausman, Jerry A., "Labor Supply," in Henry J. Aaron and Joseph A. Pechman, eds., *How Taxes Affect Economic Behavior*, Washington: The Brookings Institution, 1981.

Kay, John A., "The Deadweight Loss from a Tax System," *Journal of Public Economics*, February 1980, *13*, 111–19.

King, Mervyn A., "Welfare Analysis of Tax Reforms Using Household Data," *Journal* *of Public Economics*, July 1983, *21*, 183–214.

Pauwels, Wilfried, "Correct and Incorrect Measures of the Deadweight Loss of Taxation," *Public Finance/Finance Publiques*, 1986, *41*, 267–75.

Pazner, Elisha A. and Sadka, Efraim, "Excess-Burden and Economic Surplus as Consistent Welfare Indicators," *Public Finance/Finance Publiques*, 1980, *35*, 439–49.

Stuart, Charles, "Welfare Costs per Dollar of Additional Tax Revenue in the United States," *American Economic Review*, June 1984, *74*, 352–62.

Topham, Neville, "A Reappraisal and Recalculation of the Marginal Cost of Public Funds," *Public Finance/Finance Publiques*, 1984, *39*, 394–405.

______, "Excess Burden and the Marginal Cost of Public Spending," *Economic Letters*, 1985, *17*, 145–48.

Triest, Robert K., (1987a) "The Effect of Income Taxation on Labor Supply," unpublished doctoral dissertation, University of Wisconsin-Madison, 1987.

______, (1987b) "The Relationship Between the Marginal Cost of Public Funds and Marginal Excess Burden," Working Papers in Economics No. 194, Johns Hopkins University, 1987.

Varian, Hal R., *Microeconomic Analysis*, 2nd ed., New York: Norton, 1984.

Wildasin, David E., "Public Good Provision with Optimal and Non-Optimal Commodity Taxation: The Single-Consumer Case," *Economic Letters*, 1979, *4*, 59–64.

Effects of Spatial Price Discrimination on Output, Welfare, and Location

By HONG HWANG AND CHAO-CHENG MAI*

It is well recognized that the problem of price discrimination must address itself to the question of whether total output is greater or less under discriminatory monopoly than under simple monopoly. If discrimination does not increase output, such pricing is undesirable from the standpoint of social welfare even though it proves to be more profitable for the monopolist. To answer this question, nearly 50 years ago, Joan Robinson (1933) advanced the criterion of "adjusted concavity" to decide whether price discrimination between two submarkets would increase or decrease total output if a monopolist were seeking maximum profits. Since then, a number of writers have extended and further revised her analysis, including Edgar Edwards (1950), Eugene Silberberg (1970), Melvin Greenhut and Hiroshi Ohta (1976), James Smith and John Formby (1981), Richard Schmalensee (1981), John Formby et al. (1983), Hal Varian (1985), Jun-ji Shih, Chao-cheng Mai, and Jung-chao Liu (1988), and Hiroshi Ohta (1988). As has been well known since 1933, when the two demand curves are linear, discrimination does not change output. But, as demonstrated by Schmalensee in a linear case and generalized by Varian, welfare is reduced by allowing discrimination.

The above analyses are examined in the context of traditional non-spatial economy in which distance costs are insignificant and negligible. In a spatial world, Greenhut and Ohta (1972) considered a case where consumers' demand curves are linear and the radius of a monopolist's market area is a *variable* and then demonstrated that a spatial monopolist who adopts a spatially discriminating price would always produce a larger output than under an f.o.b. mill price policy. William Holahan (1975) further verified that spatial price discrimination results in greater welfare than a mill price policy. In addition, Martin Beckmann (1976) analyzed the same problem by assuming that the radius of a monopolist's market area is fixed exogenously and showed that the total output under discriminatory pricing is the same as under f.o.b. mill pricing. Beckmann's result is the same as Robinson's. Ohta (1988) extended these analyses by showing, on a more disaggregate level, who gains from discrimination under various alternative demand and cost conditions. Common to these writers is their basic assumption that the firm's location is predetermined. This assumption must be relaxed, however, insofar as location is an important variable and firms choose different locations under different pricing systems.

In this paper, we will treat the location as an endogenous variable and study the effects of spatial price discrimination on output, welfare and location of a monopolist in the context of spatial economy. More specifically, we will investigate the following issues: (i) Will the Robinson output theorem and the Schmalensee welfare theorem be unscathed in a spatial economy, when the radius of the monopolist's market area is fixed while its location is a decision variable? (ii) What is the optimum location under discriminatory pricing compared with the one under f.o.b. mill pricing?

*Department of Economics, National Taiwan University, Taipei, Taiwan 10020, R.O.C. and Sun Yat-Sen Institute for Social Sciences and Philosophy, Academia Sinica, Nankang, Taipei, Taiwan 11529, R.O.C. We would like to thank two anonymous referees, Richard Arnott and H. Ohta for their very useful suggestions, which led to substantial improvement in the paper. Part of the work was done while the first author was a visiting associate professor at the economics department of Clemson University. He would like to thank the department for hospitality.

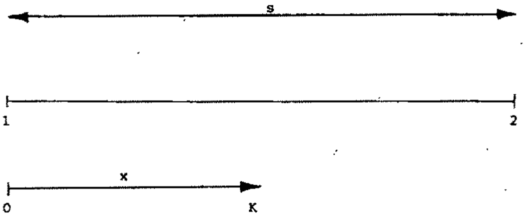


FIGURE 1. A LOCATION LINE

I. The Spatial Model with an Endogenous Location

Consider a simple setting: A spatial monopolist will locate its plant on a line in between two different markets, each of which is located at a point. It serves both markets by charging an f.o.b. mill price or discriminatory prices without fear of competition or resale between markets. The markets are s miles apart and are connected by a highway as shown in Figure 1. Let x and $(s-x)$ be the distances of its plant from markets 1 and 2, respectively.

For the purpose of this paper, we shall assume linear demand curves at each market:

$$(1)\quad q_1=\alpha-\beta p_1=\alpha-\beta(m_1+rx)$$

$$(2)\quad q_2=a-bp_2=a-b[m_2+r(s-x)],$$

where p_i, m_i, and q_i are the delivered price, mill price, and sales in market i ($i=1.2$), respectively; r is the constant transport rate; and α, β, a, and b are all positive constants. The slopes of the demand curves are measured by

$$-\frac{1}{\beta}\text{ and }-\frac{1}{b}.$$

Hence the greater the values of β and b, the flatter are the corresponding demand curves. It warrants mention that in a spatial economy, the total price paid by consumers (i.e., delivered price) differs from the net market price (net of transportation costs) received by the producer (i.e., mill price).

The demand elasticity of market i as perceived by the producer is defined as follows:

$$(3)\quad \varepsilon_1\equiv-\frac{m_1}{q_1}\frac{dq_1}{dm_1}=\frac{m_1}{\left[\frac{\alpha}{\beta}-(m_1+rx)\right]}$$

with

$$\frac{\partial\varepsilon_1}{\partial\left(\frac{\alpha}{\beta}\right)}<0\text{ and }\frac{\partial\varepsilon_1}{\partial x}>0$$

$$(4)\quad \varepsilon_2\equiv-\frac{m_2}{q_2}\frac{dq_2}{dm_2}=\frac{m_2}{\left\{\frac{a}{b}-[m_2+r(s-x)]\right\}}$$

with

$$\frac{\partial\varepsilon_2}{\partial\left(\frac{a}{b}\right)}<0\text{ and }\frac{\partial\varepsilon_2}{\partial x}<0,$$

where α/β and a/b are the price intercept values of demands.

Two related notes concerning equations (3) and (4) are warranted. First, the greater the price intercept value of demand, the less is the demand elasticity evaluated at the same price. Second, a rise in x (i.e., moving away from market 1) increases the demand elasticity in market 1 but decreases that in market 2.

In what follows, we shall use the spatial framework to derive the simple monopoly and the discrimination solutions, respectively.

A. *Simple Spatial Monopoly*

Under f.o.b. mill pricing, a spatial monopolist will quote a single price for its product. The profit function is given by

$$(5)\quad \pi^S(m,x) = [\alpha-\beta(m+rx)](m-c) + \{a-b[m+r(s-x)]\}(m-c),$$

where c denotes the constant marginal cost.

The monopolist operates on two variables to maximize its profit: the f.o.b. price, m, and the location, x. This can be accomplished by breaking the profit maximization problem down and solving it in two stages: select the optimum location, and then determine the optimal price. We proceed by maximizing profit with respect to m at any given location to obtain the comparative static effect of x on m. The first-order condition for maximization is as follows:

$$(6)\quad \frac{\partial\pi^S}{\partial m} = (\alpha+a)-\beta rx-br(s-x) + (\beta+b)c-2(\beta+b)m = 0,$$

with the second-order condition:

$$(7)\quad \frac{\partial^2\pi^S}{\partial m^2} = -2(\beta+b)<0.$$

Solving (6) yields

$$(8)\quad m^S = \frac{(\alpha+a)-\beta rx-br(s-x)+(\beta+b)c}{2(\beta+b)},$$

where m^S is the equilibrium price under simple monopoly.

From (8), we can easily derive the effect of x on m:

$$(9)\quad \frac{dm^S}{dx} = \frac{-r(\beta-b)}{2(\beta+b)} \gtrless 0 \text{ if } \beta \lessgtr b.$$

An intuitive interpretation of equation (9) follows. First, the aggregate demand curve faced by the producer is derivable as $Q = q_1+q_2=[(\alpha+a-brs)-r(\beta-b)x]-(\beta+b)m$. If the demand in market 1 is steeper (flatter) than that in market 2, that is, $\beta<b(\beta>b)$, then we see that moving the location away from market 1 increases (decreases) the quantity intercept and hence shifts out (in) the aggregate demand curve so that the optimal mill price increases (decreases).

Moreover, substituting (8) into (1) and (2) yields

$$(10a)\quad q_1^S = \frac{1}{2(\beta+b)}[(2\alpha-\beta c-2\beta rx) \times(\beta+b)-\beta(\alpha+a) + \beta^2rx+\beta br(s-x)],$$

$$(10b)\quad q_2^S = \frac{1}{2(\beta+b)} \times\{[2a-bc-2br(s-x)] \times(\beta+b)-b(\alpha+a) + \beta brx+b^2r(s-x)]\}.$$

We now turn to the problem of choosing the optimal location. Via the envelope theorem of comparative statics and noting that $(m^S-c)>0$, we can obtain[1]

$$(11)\quad \frac{d\pi^S}{dx} = \frac{\partial\pi^S}{\partial x} = -r(m^S-c)(\beta-b)\gtrless 0 \quad \text{as } \beta\lessgtr b$$

and

$$(12)\quad \frac{d^2\pi^S}{dx^2} = \frac{r^2(\beta-b)^2}{2(\beta+b)}>0.$$

[1] It is easy to demonstrate that the solutions for this problem are unchanged if we solve for a simultaneous optimization over prices and location.

Equation (12) shows that the optimized profit function is strictly convex with respect to x. It implies that any intermediate locations between markets 1 and 2 are strictly excluded in our simple location problem. Thus, combining (11) with (12), we can conclude that

$$x^S = 0 \text{ if } \beta > b, \tag{13a}$$

$$x^S = s \text{ if } \beta < b. \tag{13b}$$

Equation (13) is a very sensible outcome. As mentioned before, when $\beta < b$, increasing x shifts out aggregate demand (i.e., total sales) and hence increases the monopolist's profits. As such, the monopolist will select its optimum location at market 2 with a flatter demand. The converse is true if $\beta > b$. Thus, the slope effect of demand alone matters in location decisions under mill pricing.

From equations (10a, b) and (13a, b), we can evaluate total output under simple monopoly as follows:

$$\begin{aligned} Q^S(x=0) &= q_1^S|_{x=0} + q_2^S|_{x=0} \\ &= \frac{1}{2}[(\alpha + a) - brs \\ &\qquad - (\beta + b)c] \text{ if } \beta > b, \end{aligned} \tag{14a}$$

$$\begin{aligned} Q^S(x=s) &= q_1^S|_{x=s} + q_2^S|_{x=s} \\ &= \frac{1}{2}[(\alpha + a) - \beta rs \\ &\qquad - (\beta + b)c] \text{ if } \beta < b. \end{aligned} \tag{14b}$$

B. *Spatial Price Discrimination*

We shall now consider the case where the monopolist charges different mill prices at each market. The profit function is specified as follows:

$$\begin{aligned} \pi^D(m_1, m_2, x) &= [\alpha - \beta(m_1 + rx)](m_1 - c) \\ &\quad + \{a - b[m_2 + r(s-x)]\} \\ &\quad \times (m_2 - c). \end{aligned} \tag{15}$$

The producer will maximize its profit at any location by choosing m_i at each market to yield the first-order conditions:

$$\frac{\partial \pi^D}{\partial m_1} = \alpha - 2\beta m_1 - \beta rx + \beta c = 0, \tag{16a}$$

$$\begin{aligned} \frac{\partial \pi^D}{\partial m_2} &= a - 2bm_2 - br(s-x) \\ &\quad + bc = 0. \end{aligned} \tag{16b}$$

By solving (16a, b), we obtain the maximizing value of m_1 and m_2 for a given value of x:

$$m_1^D = \frac{1}{2}\left(\frac{\alpha}{\beta} - rx + c\right), \tag{17a}$$

$$m_2^D = \frac{1}{2}\left[\frac{a}{b} - r(s-x) - c\right]. \tag{17b}$$

The effects of a change in the producer's location on the equilibrium mill prices can be evaluated by

$$\frac{dm_1^D}{dx} = -\frac{r}{2}, \tag{18a}$$

$$\frac{dm_2^D}{dx} = \frac{r}{2}. \tag{18b}$$

It follows from (18) that under spatially discriminatory pricing, the producer will absorb one-half of the transport cost which is otherwise charged to the consumers.[2] This

[2] See, for example, the original proof in H. W. Singler (1937).

result comes from the linearity of the individual demand functions.

By substituting (17a,b) into (1) and (2), we get

$$q_1^D = \frac{\alpha - \beta r x - \beta c}{2}, \tag{19a}$$

$$q_2^D = \frac{a - br(s-x) - bc}{2}. \tag{19b}$$

Note that m_1^D, m_2^D, and the corresponding π^D are functions of x. Again by the envelope theorem of comparative statics, we have

$$\begin{aligned} \frac{d\pi^D(x)}{dx} &= \frac{\partial \pi^D}{\partial x} \\ &= r\left[b(m_2 - c) - \beta(m_1 - c)\right], \end{aligned} \tag{20}$$

$$\frac{d^2\pi^D(x)}{dx^2} = \frac{r^2}{2}(b+\beta) > 0. \tag{21}$$

Equation (21) indicates that the profit function is strictly convex with respect to x. Thus, there is no possibility of an intermediate location. Furthermore, the choice between markets 1 and 2 as the optimum location depends upon the relative magnitudes of $\pi^D(x=0)$ and $\pi^D(x=s)$ whose values are derived as follows:

$$\begin{aligned} \pi^D(x=0) &= \frac{1}{4\beta}(\alpha - \beta c)^2 \\ &\quad + \frac{1}{4b}(a - brs - bc)^2, \end{aligned} \tag{22}$$

$$\begin{aligned} \pi^D(x=s) &= \frac{1}{4\beta}(\alpha - \beta rs - \beta c)^2 \\ &\quad + \frac{1}{4b}(a - bc)^2. \end{aligned} \tag{23}$$

From equations (22) and (23), we can calculate the difference of the two profit levels:

$$\begin{aligned} \Delta &= \pi^D(x=0) - \pi^D(x=s) \\ &= \frac{rs}{2}\left[(\alpha - a) + \left(c + \frac{rs}{2}\right)(b - \beta)\right]. \end{aligned} \tag{24}$$

The sign of Δ is in general ambiguous as it depends on the two effects: (i) the quantity intercept effect, $(\alpha - a)$, and (ii) the slope effect, $(b - \beta)$. To signify the effect of an endogenous location on the output levels under the two pricing schemes, we shall assume in what follows that α and a are equal (i.e., $\alpha = a$). The case in which $\alpha \neq a$ is explored in the Appendix. Under such a circumstance, equation (24) reduces to

$$\Delta = \frac{rs}{2}\left(c + \frac{rs}{2}\right)(b - \beta) \gtrless 0 \quad \text{if } \beta \lessgtr b. \tag{25}$$

It immediately follows from (25) that

$$x^D = 0 \text{ if } \beta < b, \tag{26a}$$

$$x^D = s \text{ if } \beta > b. \tag{26b}$$

By substituting (17a,b) into (1) and (2) and noting (26a,b), we get total output under price discrimination:

$$\begin{aligned} Q^D(x=0) &= q_1^D|_{x=0} + q_2^D|_{x=0} \\ &= \frac{1}{2}\left[(\alpha + a) - brs - c(\beta + b)\right], \end{aligned} \tag{27a}$$

$$\begin{aligned} Q^D(x=s) &= q_1^D|_{x=s} + q_2^D|_{x=s} \\ &= \frac{1}{2}\left[(\alpha + a) - \beta rs - c(\beta + b)\right]. \end{aligned} \tag{27b}$$

II. Comparisons

We are now in a position to make comparisons of the optimum location, output and welfare under the two pricing policies. First of all, let us consider the optimum location. From equations (13a, b) and (26a, b), we can conclude that

$$(28a)\qquad x^S = 0 \quad \text{and}$$

$$x^D = s \text{ if } \beta > b,$$

$$(28b)\qquad x^S = s \quad \text{and}$$

$$x^D = 0 \text{ if } \beta < b.$$

This leads to the following proposition:

PROPOSITION 1: *The optimum locations under both pricing policies are corner solutions. Assuming that the two linear demand curves have an equal quantity intercept, the monopolist locates at the market with a flatter demand under f.o.b. pricing while at the market with a steeper demand under discriminatory pricing.*

This is a very striking result. The location pulls under the two pricing systems are so diverse that the patterns of the optimum locations are completely different.

Next, the comparison of output effects under f.o.b. pricing and discriminatory pricing will be made by observing the following two cases. Consider first the case of $b > \beta$, where $x^S = s$ and $x^D = 0$. Using equations (14a, b) and (27a, b), we can evaluate

$$(29a)\qquad Q^D(x=0) - Q^S(x=s) = \frac{1}{2}[rs(\beta - b)] < 0.$$

Similarly, the other case of $b < \beta$, in which $x^S = 0$ and $x^D = s$, yields

$$(29b)\qquad Q^D(x=s) - Q^S(x=0) = \frac{1}{2}[rs(b - \beta)] < 0.$$

From (29a, b), we establish the following proposition:

PROPOSITION 2: *When the monopolist's location is treated as an endogenous variable and the two linear demand curves have the same quantity intercept, total output under discriminatory pricing is definitely less than that under f.o.b. pricing.*

This outcome is sharply different from that obtained in the existing literature. In spaceless economics, Robinson (1933) and Silberberg (1970) demonstrate that total output is unchanged by discrimination when demand curves are linear. In the spatial economy, with predetermined location and fixed market radius, Beckmann (1976) obtains the same result.[3] On the other hand, Greenhut and Ohta (1972) assume that market radius is variable and the firm's location is given exogenously to show that total output is greater under spatial discrimination than under simple spatial monopoly, which is contrary to ours.

Finally, let us turn to the comparison of welfare effects under the two pricing policies. Within a nonspatial framework, Schmalensee (1981) has demonstrated that in a linear demand case, although total output remains unchanged, Marshallian welfare (consumers' surplus plus producer's profit) is reduced by allowing discrimination. Apparently, it is important to see whether this result can be applied to the spatial economy. First of all, using the demand functions (1) and (2), we can compute the consumers' surplus at each market as follows:

$$(30a)\qquad CS_1 = \frac{1}{2}\left(\frac{\alpha}{\beta} - p_1\right)q_1 = \frac{1}{2\beta}q_1^2,$$

$$(30b)\qquad CS_2 = \frac{1}{2}\left(\frac{a}{b} - p_2\right)q_2 = \frac{1}{2b}q_2^2.$$

Next, the producer's profit accruing from market i is defined as

$$(31)\qquad \pi_i = (m_i - c)q_i \quad \text{for} \quad i = 1, 2.$$

Using equations (30a, b) and (31), the welfare under simple spatial monopoly is deriv-

[3] For any given x, it can be easily shown that $Q^D = Q^S$.

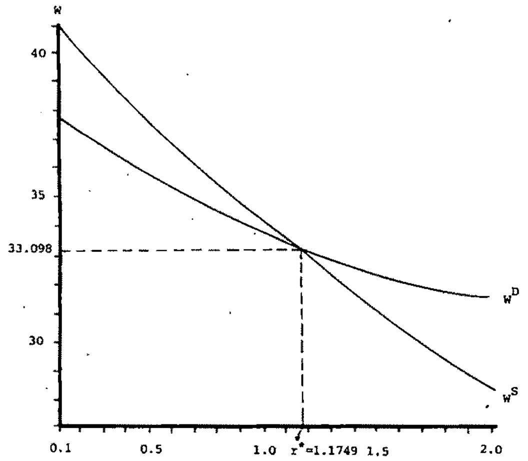


FIGURE 2. WELFARE COMPARISON

able as

$$(32)\quad W^S = CS_1^S + CS_2^S + \pi_1^S + \pi_2^S,$$

$$= \frac{1}{2\beta b(\beta + b)}$$

$$\times\left\{(\beta + b)\left[b(q_1^S)^2 + \beta(q_2^S)^2\right]\right.$$

$$\left.+ 2\beta b(q_1^S + q_2^S)^2\right\},$$

where q_1^S and q_2^S are defined in (10a,b).

Similarly, the welfare under spatial price discrimination is given by

$$(33)\quad W^D = CS_1^D + CS_2^D + \pi_1^D + \pi_2^D,$$

$$= \frac{3}{2}\left[\frac{1}{\beta}(q_1^D)^2 + \frac{1}{b}(q_2^D)^2\right],$$

where q_1^D and q_2^D are defined in (19a,b).

Comparing (32) and (33), we obtain

$$(34)\quad W^D - W^S \gtrless 0,$$

as it depends on particular parameter values of demand, cost, and transport rate structures. For example, let $s = c = \beta = 1$, $a = \alpha = 10$ and $b = 2.6$, the relation in (34) can be graphed in Figure 2 in terms of transport rate structure. It shows that $W^D \gtrless W^S$ if $r \gtrless r^* = 1.1749$.

PROPOSITION 3: *When the firm's location is treated as an endogenous variable, the welfare level under spatial price discrimination could be greater than that under simple monopoly even if both demand curves are linear.*

This result weakens Schmalensee's conclusion derived from a nonspatial framework that f.o.b. mill pricing is always preferred. It is also different from Holahan's (1975) claim derived from a spatial model that the welfare under discriminatory pricing is greater than that under mill pricing. This difference is due to his assumption that the monopolist's location is fixed regardless of pricing structures.

III. Conclusions

This paper has endeavored to examine the effects of spatial price discrimination on output, welfare, and location of a monopolist. The main conclusions are provided as follows:

(1) When the two demand curves are linear, there is no possibility of an intermediate location. The monopolist will locate at the site of the market with steeper demand under discriminatory pricing, while at that of the market with flatter demand under f.o.b. pricing.

(2) When the location is treated as a decision variable and both demand curves are linear with the same quantity intercept values, then the total output under discriminatory pricing is less than that under mill pricing, but the welfare effect of spatial price discrimination is indeterminate. These results are sharply different from the ones derived in previous studies.

Although our analysis may be too simple to draw the conclusion that price discrimination should be prohibited, it has at least demonstrated an example in a spatial context that questions the blanket prohibition of price discrimination.

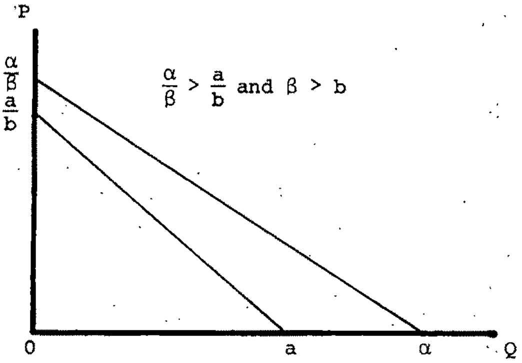


FIGURE A1. $\Delta > 0$

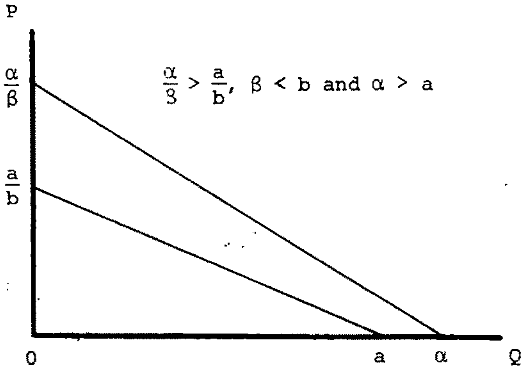


FIGURE A2. $\Delta > 0$

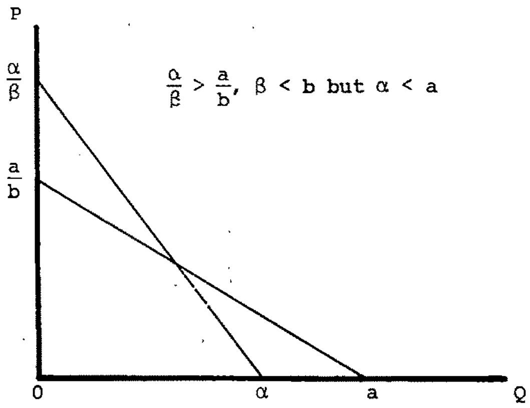


FIGURE A3. $\Delta \gtrless 0$

APPENDIX

This appendix discusses the optimum location under discriminatory pricing when the two quantity intercept values are not equal.

Let us first rewrite equation (24) in the text as follows:

$$\text{(A1)}\quad \Delta \equiv \pi^D(x=0)-\pi^D(x=s)$$
$$=\frac{rs}{2}\left[(\alpha-a)+\left(c+\frac{rs}{2}\right)(b-\beta)\right]\gtrless 0.$$

If $\Delta > 0$, then it implies that the firm will locate at market 1 (i.e., $x=0$). The converse is true if $\Delta < 0$. In what follows, let us consider the case of $\Delta > 0$. Under such a circumstance, (A1) may be rearranged in the following form:

$$\text{(A2)}\qquad \frac{\alpha}{\beta}>\left(\frac{b}{\beta}\right)\frac{a}{b}+\left(1-\frac{b}{\beta}\right)\left(c+\frac{rs}{2}\right).$$

Note that $(c+rs)<\min[\alpha/\beta, a/b]$ for both markets to be served under conditions of extreme location either at $x=0$ or $x=s$. With this note in mind, we now consider the following two sets of conditions, both of which yield $\Delta > 0$.

Case (*i*): $\alpha/\beta > a/b$ and $\beta > b$.

This case means that the less elastic demand (or equivalently, the demand with a smaller price intercept) is flatter than the more elastic demand as pictured in Figure A1. By noting $a/b > (c+rs/2)$, this set of conditions guarantees condition (A2) above.

Case (*ii*): $\alpha \geq a$ and $\beta < b$.

The set of conditions directly establishes $\Delta > 0$ via (A1). These conditions also imply $\alpha/\beta > a/b$ as shown in Figure A2. This states that even if the less elastic demand is steeper than the more elastic demand (i.e., $\alpha/\beta > a/b$ and $\beta > b$), the sign of Δ is still positive *provided* that $\alpha \geq a$. We then conclude that the demand elasticity alone matters in location decisions under discriminatory pricing, provided $\alpha \geq a$. This proviso requires that the less elastic demand be *not* smaller than the more elastic demand for any price $P < a/b$. In contrast to the elasticity criterion, this proviso refers to the quantity intercept values and therefore may be termed as the (demand) volume criterion, which requires the monopolist to locate in the market with greater volume of demand.

Combining cases (*i*) and (*ii*), we may conclude that as long as this proviso holds, the elasticity criterion is *sufficient* to yield $\Delta > 0$ regardless of the demand slope (i.e., no matter whether $\beta > b$ or $\beta < b$), that is, discriminatory monopolist will locate at the market with less elastic demand if this proviso holds. However, the sign of Δ is not definite if the proviso is violated as drawn in Figure A3.

REFERENCES

Beckmann, Martin J., "Spatial Price Policies Revisited," *Bell Journal of Economics*, Autumn 1976, 7, 619–30.

Edwards, Edgar O., "The Analysis of Output

Under Discrimination," *Econometrica*, April 1950, *18*, 163–72.

Formby, John P., Layson, Stephen K. and Smith, W. James, "Price Discrimination, 'Adjusted Concavity,' and Output Changes Under Conditions of Constant Elasticity," *Economic Journal*, December 1983, *93*, 892–99.

Greenhut, Melvin L. and Ohta, Hiroshi, "Output Under Alternative Spatial Pricing Techniques," *American Economic Review*, September 1972, *62*, 705–13.

______ **and** ______, "Joan Robinson's Criterion for Deciding Whether Market Discrimination Reduces Output," *Economic Journal*, March 1976, *86*, 96–97.

Holahan, William L., "The Welfare Effects of Spatial Price Discrimination," *American Economic Review*, June 1975, *65*, 498–503.

Ohta, Hiroshi, *Spatial Price Theory of Imperfect Competition*, College Station: Texas A&M University Press, 1988.

Robinson, Joan, *The Economics of Imperfect Competition*, London: Macmillan, 1933.

Schmalensee, Richard, "Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination," *American Economic Review*, March 1981, *71*, 242–47.

Shih, Jun-ji, Mai, Chao-cheng and Liu, Jung-chao, "A General Analysis of the Output Effect Under Third-Degree Price Discrimination," *Economic Journal*, March 1988, *98*, 149–58.

Silberberg, Eugene, "Output Under Discriminatory Monopoly," *Southern Economic Journal*, July 1970, *37*, 84–87.

Singler, H. W., "A Note on Spatial Price Discrimination," *Review of Economic Studies*, June 1937, *5*, 75–77.

Smith, W. James and Formby, John P., "Output Changes Under Third-Degree Price Discrimination: A Reexamination," *Southern Economic Journal*, July 1981, *48*, 164–71.

Varian, Hal R., "Price Discrimination and Social Welfare," *American Economic Review*, September 1985, *75*, 870–75.

On the Equilibrium Yen-Dollar Rate

By HIROSHI YOSHIKAWA*

This paper presents a definition of the long-run equilibrium exchange rate that emphasizes the role of supply factors in addition to the more traditional price differential variables. Using this definition, we estimate the long-run equilibrium yen-dollar rate during the period of 1973–87. This calculation demonstrates that supply factors are in fact very important determinants of the long-run movement of the exchange rate.

Despite the central role the exchange rates play in the economy, economists seem to doubt their own ability to explain, let alone predict, exchange rate movements. The reason is that in many cases structural exchange rate equations do not significantly outperform various other models that do not appeal to economic theory such as the random walk. (See, for example, Richard Meese and Kenneth Rogoff, 1983.) While we also share this pessimism concerning our ability to forecast exchange rates in the short run (intradaily to monthly or quarterly), we argue that it is possible to reasonably *explain*, if not predict, long-run movements.

The long-run equilibrium rate is most often thought to be determined by purchasing power parity (PPP), which tends to focus only on price differentials. Medium-run divergences of the exchange rate from the long-run equilibrium are therefore commonly identified with fluctuations in the inflation-adjusted real exchange rate. This paper demonstrates, however, that the *simple* PPP is unsatisfactory because it emphasizes only the inflation differential. In the more elaborated theory of purchasing power parity, of course, the potential importance of *supply-side real factors* as critical determinants of the exchange rate has been well recognized. (See, for example, John M. Keynes, 1923; Rüdiger Dornbusch, Stanley Fischer, and Paul A. Samuelson, 1977; Alan Stockman, 1980.) Nevertheless, in empirical applications and policy discussions these *supply-side real factors* have been curiously ignored: To our knowledge, an exception is David Hsieh (1982). The main point of this paper is to show that these supply-side factors actually have been very important determinants of the yen-dollar rate over the period of 1973 to 1987.

I. The Definition of the Equilibrium Rate

In this section, we present a definition of the long-run equilibrium exchange rate. We start by considering the tradables sector of the Japanese economy. We assume that two inputs, labor and natural resources (including energy), are necessary for production. For simplicity, we also assume a fixed coefficient production function. Accordingly, the production process can be characterized by two parameters, a and b. Here a and b are, respectively, the labor input and the natural resources input necessary to produce one unit of the tradable good.

Under the assumption of perfect competition, we are led to the following relationship:

$$wa + p_R b = p. \qquad (1)$$

Here p and p_R are the domestic (yen) prices of tradables and natural resources, respectively, and w is the nominal wage rate. For simplicity, we have assumed zero profits in (1), although this assumption is not essential to our analysis. If we allow for a *normal* rate of profit and call it r, we need only to add rp to the left-hand side of equation (1).

*Faculty of Economics, University of Tokyo, Tokyo, Japan. I am indebted to Professors Tatsuo Hatta, Takatoshi Itoh, Masahiro Kuroda, Kenichi Ohno, Kunio Okina, Yoshiyasu Ono, Kazuo Sato, Kazuyuki Suzuki, Kazuo Ueda, and two anonymous referees for their helpful comments and suggestions. I also acknowledge with gratitude the excellent research assistance of Jennifer Holt Dwyer, Tohru Inoue, and Elizabeth Norville. The research reported here was supported by the Kikawada Foundation.

As long as we hold the normal rate of profit r constant, this inclusion will not affect our analysis.

Assume that Japan exports the tradable goods and imports the natural resources that are necessary for the production of the tradables. In 1980, for example, 96 percent of Japan's exports consisted of manufactured products, whereas 70 percent of imports were raw materials including oil. We take as given, in dollars, both p^*, the international price of the tradable good, and p_R^*, the international price of the natural resources. Since most manufactured products are more or less differentiated, the "law of one price" does not exactly hold. We would therefore expect the following relation between p and p^*:

$$p = ep^*/\varepsilon. \tag{2}$$

Here e (yen/dollar) is the nominal exchange rate, and ε is random deviations due to such factors as changes in the relative demand for Japan's exportables. We assume that the expected value of ε, $E(\varepsilon)$ is one. For the natural resources, we have

$$p_R = ep_R^*. \tag{3}$$

With (2) and (3), (1) can be rewritten as

$$wa + ep_R^*b = ep^*/\varepsilon. \tag{4}$$

If we set p^* at the U.S. unit costs, we have

$$\begin{aligned} p^* &= w^*a^* + p_R^*b^* \\ &= w^*\left[a^* + \left(\frac{p_R^*}{w^*}\right)b^*\right], \end{aligned} \tag{5}$$

where a^* and b^* are the labor and energy coefficients and w^* is the nominal wage rate in the United States. When (5) is substituted into (4), we obtain

$$\begin{aligned} e = {} & \varepsilon(w/w^*) \\ & \times[a/(1-bt^*)]\Big/\left[a^* + \left(\frac{p_R^*}{w^*}\right)b^*\right], \end{aligned} \tag{6}$$

where $t^* = \varepsilon p_R^*/p^*$ is the relative price of natural resources in terms of Japan's exportables. Since $E(\varepsilon) = 1$, we finally obtain

$$\begin{aligned} E(e) = {} & (w/w^*) \\ & \times[a/(1-bt^*)]\Big/\left[a^* + \left(\frac{p_R^*}{w^*}\right)b^*\right] \end{aligned} \tag{7}$$

for a long-run trend of yen-dollar rate. (7) defines the long-run equilibrium exchange rate. The equilibrium exchange rate depends on the following factors:

(i) The relative nominal wage level in Japan and in the United States
(ii) The terms of trade (the relative price of natural resources, such as oil, and of manufactured goods in the international market)
(iii) The technological parameters of the Japanese export sector.

The simple PPP tends to emphasize the price differential, which in this case is factor (i) or the nominal wage differential. In contrast to factor (i), factor (ii), the terms of trade, and factor (iii), the productivities, are real factors.

At this point, it is appropriate to clarify the meaning of the above equilibrium exchange rate. The long-run equilibrium rate defined by (7) is the exchange rate at which the profit rate of the Japanese export sector remains constant, given exogenous movements of wages, productivity, terms of trade and export prices. It is certainly compatible with a variety of different models.[1] This definition of long-run equilibrium rate does not, however, take into account capital accumulation and intertemporal saving behavior or current account imbalances. In this sense, it is simply one version of PPP by which we can measure the relative importance of monetary and real factors in determining the long-run movement of the exchange rate.

Given this equilibrium rate, we observe the following relationship between the do-

[1] A classical model that produces the equilibrium rate defined by (7) is in Yoshikawa (1987).

mestic and foreign prices of the *tradables*:

$$(8) \qquad E(e)p^*/p = 1.$$

This means that in order to hold constant the competitiveness of both foreign and domestic countries with regards to the tradables sector, the p adjusted real exchange rate must be one.

However, the real exchange rate calculation is often based on more broadly defined price deflators such as the GNP deflator or the CPI. In such instances, if we let $\hat{p}$, $\hat{p}^*$ represent the domestic and foreign deflators, respectively, (8) is violated for $\hat{e}$, even though (9) holds:

$$(9) \qquad \hat{e}\hat{p}^*/\hat{p} = 1.$$

To be more specific, the GNP deflators of home (Japan) and foreign (the U.S.) countries are respectively defined as

$$(10) \qquad \hat{p} = \delta p + (1-\delta)p_N$$

and

$$(11) \qquad \hat{p}^* = \delta^* p^* + (1-\delta^*)p_N^*,$$

where p_N and p_N^* are the domestic and foreign prices of nontradables, and δ and δ^* are the expenditure shares of tradables in Japan and the United States, respectively.

Under the condition

$$(12) \qquad \delta^* + (1-\delta^*)\left(\frac{p_N^*}{p^*}\right) < \delta + (1-\delta)\left(\frac{p_N}{p}\right),$$

if (9) holds, we have

$$(13) \qquad p < \hat{e}p^*.$$

If $\delta^* = \delta$, (12) boils down to

$$(14) \qquad \frac{p}{p_N} < \frac{p^*}{p_N^*}.$$

Therefore, if high productivity in Japan makes tradables *relatively* cheaper in Japan than in the United States, Japan's competitiveness is increased even though the real exchange rate in terms of the GNP deflator remains constant. This point is also emphasized by Richard C. Marston (1986) and Ken-ichi Ohno (1986).

II. The Estimation of the Equilibrium Rate

In this section, we estimate the equilibrium yen-dollar rate on the basis of (7). We adopt the two-country framework, using Japan as the home country and the United States as the foreign country. Our analysis is limited in the sense that we take into account only one trading partner. Based on this assumption, we calculate (7).

First, a and b are calculated from the input coefficients (annual data) of thirteen manufacturing industries as categorized by SNA, which are used in the Economic Planning Agency's medium-term multisector model. We identify Japan's export industries as fibers, chemicals, primary metals, general machinery, electrical equipment, transportation equipment, and precision equipment ($i = 1,2,\ldots,7$). Using the export volume of each industry, X_i, ($i = 1,2,\ldots,7$) the weights are calculated:

$$(15) \qquad \sigma_i = X_i/\Sigma_1^7 X_j \quad (i = 1,2,\ldots,7).$$

Here X_i refers to exports (in dollar terms) after passing customs. Data are taken from the Ministry of Finance's *Foreign Trade Outlook*. Using these weights,

$$(16) \qquad a \equiv \Sigma_1^7 \sigma_i a^i,$$

where a^i is each industry's labor input coefficient.

Similarly b is calculated as follows:

$$(17) \qquad b \equiv \Sigma_1^7 \sigma_i [b_{i0} + b_{i,PETRO} b_{PETRO,0}],$$

where b_{i0} is the ith industry's input coefficient for oil and natural gas. Here $b_{i,PETRO}$ and $b_{PETRO,0}$ are respectively the ith industry's input coefficient for petroleum products and the petroleum products sector's input

TABLE 1—LABOR AND ENERGY COEFFICIENTS IN JAPAN'S EXPORT INDUSTRIES

	Labor Coefficient	Energy Coefficient
1973	0.089	0.026
1974	0.081	0.030
1975	0.087	0.024
1976	0.079	0.020
1977	0.074	0.018
1978	0.067	0.016
1979	0.061	0.017
1980	0.058	0.017
1981	0.056	0.012
1982	0.054	0.010
1983	0.053	0.009
1984	0.049	0.008
1985	0.049	0.008
1986	0.044	0.006
1987	0.041	0.004

coefficient for oil and natural gas. Using this method, a and b are prepared annually. The annual series for 1973–87 are shown in Table 1. We observe that in the 10 years-plus period after the first oil shock, the labor coefficient and the energy-natural resources coefficient of Japan's exporting industry were down to about one-half and one-sixth, respectively.

Next, t^* was calculated from the contract-based (mostly dollar-based) export price of manufactured goods and the import prices of oil, gas, and coal. The data source for this calculation is the Bank of Japan's *Annual Statistics of Prices*. Here w is the nominal wage in Japan's manufacturing industry, and is taken from the Bank of Japan's *Monthly Economic Statistics*.

Finally, to obtain the equilibrium rate defined by (7), we must measure the U.S. unit costs of the products that correspond to Japan's exportables. To calculate these unit costs, we start with the following definition:

$$\text{(18)} \quad \text{Unit Cost} = \frac{w^*L^* + p_M^*M^*}{Q^*},$$

where w^* and p_M^* are the nominal wage and price of materials and energy. L^*, M^*, and Q^* are the U.S. labor, material inputs and output, respectively. The *Annual Survey of Manufacturers*; *Statistics for Industry Groups and Industries* compiled by the U.S. Department of Commerce provides the necessary data. Using these data, we calculate the unit cost as follows:

$$\text{(19)} \quad \text{Unit Cost} = p^* \times \left(\frac{w^*L^*}{P^*Q^*} \right) + p^* \times \left(\frac{p_M^*M^*}{P^*Q^*} \right).$$

In this equation, w^*L^*, $p_M^*M^*$, and P^*Q^* are, respectively, the payroll of all employees, the cost of materials, and the sum of the cost of materials and value added by manufactures; p^* in (19) is WPI. We prepared this unit cost on an annual basis for the above-mentioned seven industries, which correspond to Japan's export industries, and then with weights (15) calculated the economywide unit cost. This is the unit cost $w^*a^* + P_R^*b^*$ in (7).

To calculate (7), we need a base year. Here we consider two cases: The first case presumes that the actual exchange rate was close to (7) prior to the first oil shock (the first quarter of 1973), and the second case takes 1975 when Japan's current account was nearly in balance.

The estimated long-run equilibrium rates are shown in Table 2. They correspond to two different base years. For the sake of comparison, we also show simple PPP rates based on *economywide* wages that are supposed to reflect general inflation differential but do not take into account any productivity differential. Finally, we show the actual yen-dollar exchange rate. They are all shown in Figures 1 and 2.

Our equilibrium rate differs substantially from simple PPP rate based on only *economywide* wages. Since our equilibrium rate broadly tracks the trend of the actual exchange rate, we argue that the supply-side factors such as labor and energy productivity are very important determinants of the exchange rate in the long run. The results obtained are reasonably robust with respect to changes in the base year.

We can measure the relative impact of each factor on the equilibrium rate by calculating the hypothetical rate. In order to do

TABLE 2—WAGE-ADJUSTED RATES AND EQUILIBRIUM EXCHANGE RATES

		Base Year 1973		Base Year 1975	
	Actual Rate	Equilibrium Rate	Wage-Adjusted Rate	Equilibrium Rate	Wage-Adjusted Rate
1973	271.70	271.70	271.70	284.37	228.30
1974	292.08	271.79	324.83	284.46	272.95
1975	296.79	283.56	353.21	296.79	296.79
1976	269.55	278.10	370.25	291.07	311.11
1977	268.51	269.02	373.39	281.57	313.75
1978	210.44	240.43	367.82	251.64	309.08
1979	219.14	217.15	361.96	227.28	304.15
1980	226.74	199.70	359.84	209.01	302.37
1981	220.54	182.93	348.47	191.46	292.81
1982	249.08	176.63	347.12	184.87	291.68
1983	237.51	175.68	339.78	183.87	285.51
1984	237.52	166.04	337.53	173.79	283.62
1985	238.54	156.58	339.27	163.89	285.08
1986	168.52	148.13	342.18	155.04	287.52
1987	144.64	140.53	339.95	147.08	285.65

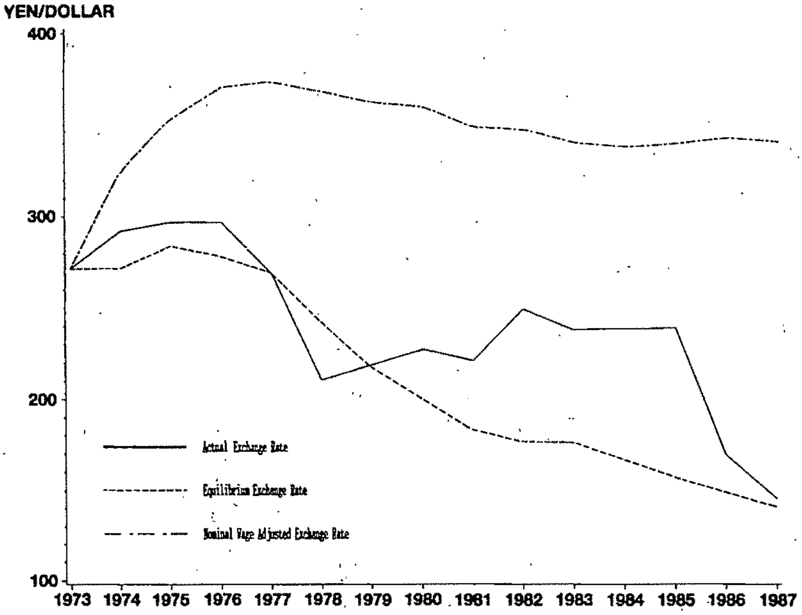


FIGURE 1. EQUILIBRIUM AND ACTUAL EXCHANGE RATES (BASE YEAR, 1973)

so, we assume that one factor remains constant at the base-year level. These hypothetical rates for the constant a, b, the U.S. unit labor cost and energy cost, are shown together with the equilibrium rate for the 1975 base-year case in Table 3. We can see, for example, that the equilibrium rate would have been 303.77 yen per dollar in 1987 if the labor coefficient in Japan had remained unchanged at the 1975 level. As we can see in Table 1, in reality, the labor coefficient in Japan declined sharply between 1975 and 1987, and therefore the equilibrium rate was 147.08 yen per dollar. The divergence between the equilibrium rate and the hypothetical rate for the constant b, the Japanese energy coefficient, is rather small, which means that the effect of changes in b on the equilibrium rate is not substantial. Thus, overall, an increase in labor productivity in

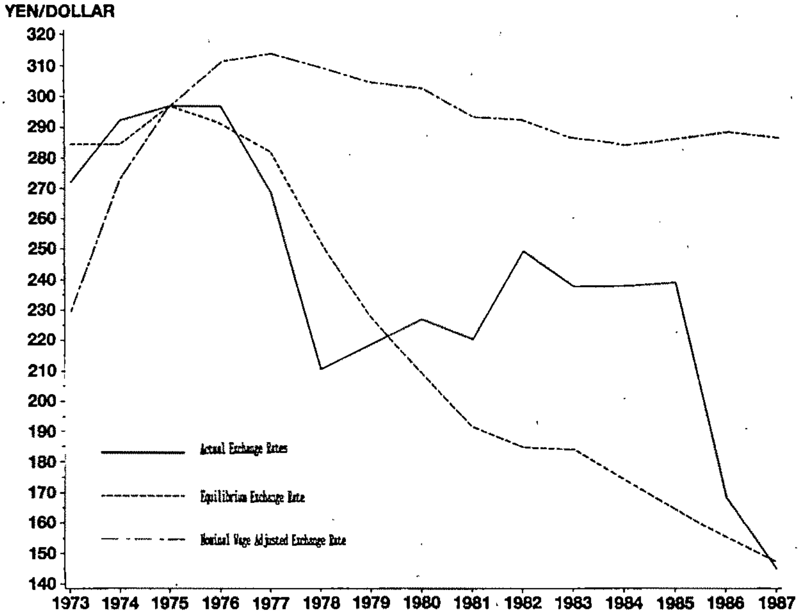


FIGURE 2. EQUILIBRIUM AND ACTUAL EXCHANGE RATES (BASE YEAR, 1975)

TABLE 3—THE HYPOTHETICAL EQUILIBRIUM RATE

	Equilibrium Rate	Constant Japanese Labor Coefficient	Constant Japanese Energy Coefficient	Constant U.S. Unit Labor Cost	Constant U.S. Unit Material Cost
1973	284.37	277.20	283.97	266.30	239.77
1974	284.46	305.88	283.45	276.36	260.91
1975	296.79	296.79	296.79	296.79	296.79
1976	291.07	321.82	291.72	291.21	302.67
1977	281.57	331.04	282.56	283.68	305.75
1978	251.64	324.64	252.66	257.35	285.27
1979	227.28	323.48	228.37	235.92	271.47
1980	209.01	311.55	210.76	223.26	265.90
1981	191.46	297.84	194.34	208.09	257.47
1982	184.87	297.64	188.02	204.73	250.41
1983	183.87	300.94	187.09	202.75	253.93
1984	173.79	308.33	176.73	189.43	246.44
1985	163.89	308.53	166.89	180.23	234.65
1986	155.04	306.21	156.87	171.04	222.42
1987	147.08	303.77	148.88	163.14	213.19

Japan is by far the most important factor behind the long-run appreciation of the yen.[2] Conversely, on the U.S. side, an increase in unit energy and material costs seems to be a more important determinant of the dollar's depreciation than is an increase in the unit labor cost.

[2] As John R. Norsworthy and David H. Malmquist (1983) show, the substantial increase in labor productivity in Japan can be attributed to the concentration of capital investment in Japan's manufacturing sector, particularly in the machinery industries. This high concentration of investment since the first oil shock of 1973 has led to the creation of what can be called the "machinery block." See Hiroshi Yoshikawa and Fumio Ohtake (1987).

III. Conclusion

In this paper, we have proposed a definition of the long-run equilibrium yen-dollar rate and attempted to estimate it by using Japanese and U.S. data. The long-run equilibrium rate defined in Section I is based on three factors: (1) the (wage) inflation differential between Japan and the United States; (2) the *real* world price of natural resources, such as oil, in terms of manufactured goods (Japan's terms of trade); and (3) the labor and energy coefficients in both Japan and the United States for Japan's export sector. As one variation of PPP it allows us to explore the relative importance of monetary and real factors in determining the long-run trend of the exchange rate.

From 1973 to 1987, the trend in the actual exchange rate broadly coincides with the movement of our equilibrium rate. On occasion, however, the actual exchange rate diverges from its long-run equilibrium value for fairly long periods. For example, the yen was greatly overvalued in 1978 and the dollar was overvalued for more than four years, from 1981 through 1985. During most of the period of the dollar's overvaluation, from 1981 through 1984, the equilibrium yen rate continued to appreciate but the actual yen rate depreciated. This divergence can only be explained by fiscal and monetary policies. The depreciation of the dollar from 1985, on the other hand, was a return to the equilibrium value.

Our analysis demonstrates that real factors, most notably, changes in labor productivity, have been much more important than the nominal wage differential in determining the long-run trend of the yen-dollar rate from 1973 through 1987. As mentioned above, most empirical works concerning the long-run exchange rate tend to focus on price differential and have largely ignored productivity. Meanwhile, statistical evidence indicates that changes in both nominal and real exchange rates can be approximated by the random walk, and therefore that they tend to be nearly permanent. Because changes in real and nominal exchange rates are very highly correlated and have similar variances, it is consistent with the view that most changes in *nominal* exchange rates are due to *real* shocks with a large permanent component. Purely statistical analyses leading to this conclusion, however, cannot explain what kind of real shocks are in fact important. This paper has identified increases in labor productivity in Japan's export industries as the most important factor to explain the long-run trend of the yen-dollar rate from 1973 through 1987. Given our findings, we hope that productivity will be given greater weight in future analysis of long-term trends of exchange rates.

REFERENCES

Dornbusch, Rudiger, *Dollars, Debts, and Deficits*, Cambridge, MA: MIT Press, 1986.

______, Fischer, Stanley and Samuelson, Paul A., "Comparative Advantage, Trade and Payments in a Ricardian Model with a Continuum of Goods," *American Economic Review*, December 1977, *67*, 823–39.

Hsieh, David, "The Determination of the Real Exchange Rate," *Journal of International Economics*, May 1982, *12*, 355–62.

Keynes, John M., *A Tract on Monetary Reform*, London: Macmillan, 1923.

Krugman, Paul, "Is the Strong Dollar Sustainable?" in *The U.S. Dollar: Prospects and Policy Options*, Kansas City: Federal Reserve Bank of Kansas City, 1985.

Marston, Richard C., "Real Exchange Rates and Productivity Growth in the United States and Japan," NBER Working Paper No. 1922, May 1986; forthcoming in J. D. Richardson and S. Arndt, eds., *Real Financial Linkages in Open Economy*, Cambridge, MA: MIT Press.

Meese, Richard and Rogoff, Kenneth, "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?" *Journal of International Economics*, February 1983, *14*, 3–24.

Norsworthy, John R. and Malmquist, David H, "Input Measurement and Productivity Growth in Japanese and U.S. Manufacturing," *American Economic Review*, December 1983, *73*, 947–67.

Ohno, Ken-ichi, "Yen-Dollar Rate Misalignments: A New Purchasing Power Parity Criterion," mimeo., Stanford University, May 1986.

Stockman, Alan, "A Theory of Exchange Rate Determination," *Journal of Political Economy*, August 1980, *88*, 673–98.

Yoshikawa, Hiroshi, "On the Equilibrium Yen Dollar Rate," Discussion Paper No. 154, I.S.E.R., Osaka University, October, 1987.

______ **and Ohtake, Fumio,** "Postwar Business Cycles in Japan: A Quest for the Right Explanation," *Journal of Japanese and International Economies*, December 1987, *1*, 373–407.

Bank of Japan, *Annual Statistics of Prices*, Tokyo, 1974–1988.

______, *Monthly Economic Statistics*, Tokyo, January 1974–December 1988.

Ministry of Finance, Japan, *Foreign Trade Outlook*, Tokyo, 1974–1988.

U.S. Department of Commerce, *Annual Survey of Manufacturers; Statistics for Industry Groups and Industries*, Washington, 1974–1988.

On the Basing-Point System

By Bruce L. Benson, Melvin L. Greenhut, and George Norman*

Jacques F. Thisse and Xavier Vives observed in this journal (1988) that analysis of the base-point price (BPP) system "should consider its role as a coordinating and collusive device" (p. 12). This conclusion has much in common with that of Fritz Machlup (1952) and Clair Wilcox (1955) and provides a sharply contrasting thesis from that of David D. Haddock, who had contended in (1982) that BPP systems are competitive. However, Thisse and Vives (henceforth TV) were not specifically, or even primarily concerned with base-point pricing. Rather, they focused attention on noncooperative strategies in spatial markets, deriving several conclusions by approaching the choice of pricing policy from an explicitly strategic, game-theoretic viewpoint. While their results do imply that noncooperative base-point pricing is unlikely, they did not center attention on the conditions that must exist in order for a *competitive* base-point pricing system to arise.

The present paper extends TV's theoretical analysis by focusing exclusively on base-point pricing *and* by ascertaining the conditions that would be required for Haddock's noncooperative BPP system to arise. In particular, the initial part of this paper utilizes the TV framework to identify the necessary conditions for a noncooperative BPP system. The TV framework leads us to a different interpretation than Haddock proposed. The second part of this paper adds, however, to the contention that the system *could indeed have* competitive origins; but we shall observe that it limits such claim to certain highly restrictive conditions. This part of the paper also refers to the reality of BPP. A third and concluding section accounts for the "cooperative" use of the system when, in fact, other "collusive" delivered pricing systems that would be more profitable than BPP can be conceived of.

I. Theory

For simplicity, assume that all consumers are identical, and are distributed over a line market in which there are two production sites I and II, with site I the lower-cost site (perhaps because of better access to raw materials). Assume that there are two firms, $i = I, II$ selling a homogeneous product, with firm i located at site i. Transport costs per unit of output from site i to consumer location x are given by a nonnegative, increasing function $t_i(|i-x|)$, where $|\cdot|$ is the distance norm, and $t_i(0)=0$. Figure 1 (see also Haddock, Figure 1) illustrates such a case, with mc_i denoting marginal production cost at site i, and $m_i(x)$ denoting marginal cost of production and transportation from site i to consumer location x: $m_i(x) = mc_i + t_i(|i-x|)(i = I, II)$.

Define the monopoly price $p_i^M(x)$ that firm i would charge to consumers at x as the price that would maximize profit to firm i from consumers at x in the absence of any competition (this price is derived from the standard MR = MC condition). TV show that if the duopolists simultaneously choose their pricing policy and price, the noncooperative equilibrium pricing strategy $p_i^*(x)$ for firm i is

$$(1)\quad p_i^*(x) = \begin{cases} p_i^M(x) \text{ if } p_i^M(x) \le m_j(x) \\ m_j(x) \text{ if } p_i^M(x) > m_j(x), \\ \quad \text{and } m_i(x) \le m_j(x) \\ m_i(x) \text{ otherwise,} \end{cases}$$

*Department of Economics, Florida State University, Tallahassee, FL 32306-2045; Department of Economics, Texas A&M University, College Station, TX 77843-4228 and adjunct, University of Oklahoma, Norman, OK 73019; and Department of Economics, The University of Leicester, Leicester, England LE1 7 RH.

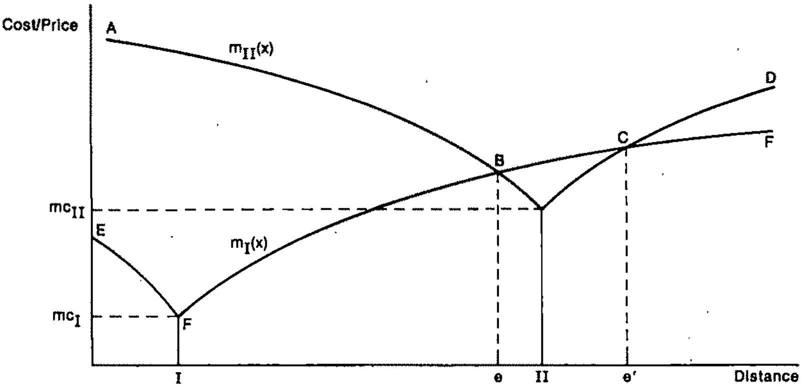


FIGURE 1

a form of price discrimination first discussed by Edgar M. Hoover (1937).[1] Assume without loss of generality that $p_i^M(x)$ everywhere exceed marginal cost of production and transportation for firm j $(j \neq i)$. Then the price equilibrium is illustrated in Figure 1 by the lines *AB*, *CD* for firm *I* and *BC* for firm *II*. If the pricing game is a two-stage game, in which firms first choose a pricing policy and then compete in prices contingent on the chosen pricing policies, TV show again that the discriminatory policy of equation (1) is the unique (Nash) equilibrium outcome.

Is pricing policy (1) consistent with BPP? It is quite clear that firm *II* is following a BPP with site *I* as base, *but firm I is also following BPP with site II as base*. In other words, the only equilibrium is a peculiar type of multiple BPP system in which consumers are charged a price less by the amount ε than the marginal costs plus freight of the second-lowest cost supplier.

Now consider what will arise if a second firm is introduced at site *I*: if, for example, the producer at site *II* has a second plant at site *I*. Using precisely the same argument, the resulting Nash equilibrium prices are now given by

$$(2) \qquad p_I^*(x) = m_I(x)$$

$$p_{II}^*(x) = \begin{Bmatrix} m_I(x) \text{ if } m_I(x) > m_{II}(x) \\ m_{II}(x) \text{ if } m_I(x) < m_{II}(x) \end{Bmatrix}.$$

These are illustrated in Figure 1 by the lines EFB and CF for the firms at site *I* and *BC* for the firm at site *II*. Pricing policy (2) is the policy derived by Haddock (although Haddock does so by requiring that production at site *I* be "competitive"): it is BPP with site *I* as base.[2]

II. Conditions for Competitive BPP

We can now identify the somewhat restrictive condition under which this policy emerges. For a single-base basing-point pricing system to be the outcome of a competitive process, it is necessary that production at the basing point be in some sense compe-

[1] This equilibrium assumes that individual demand is "well behaved," in that profit from consumers at x is quasi-concave in price.

[2] Recall that the pricing equilibria assume $p_i^M(x) > m_j(x)(j \neq i)$ for all x. Note that Haddock ignores the possibility that firm *II* (the local monopolist) could have a sufficiently strong monopoly advantage near site *II* as to be able, and so willing, to charge the monopoly price.

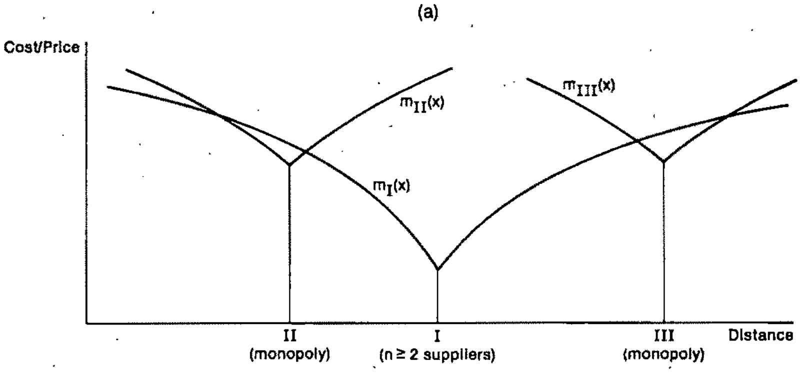


FIGURE 2. PANEL (a)

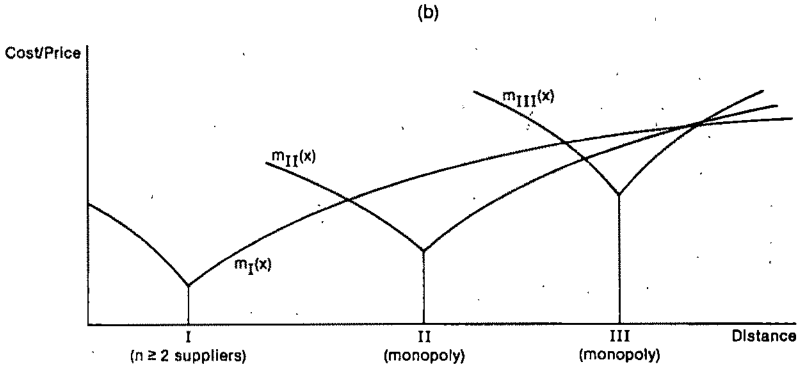


FIGURE 2. PANEL (b)

titive[3] *and production at distant sites be monopolized.* It is *further necessary that firms at the basing point* always *be either the lowest or the second-lowest cost suppliers to all consumers.* Extending the analysis, a single-base BPP system will also arise from the configuration of suppliers in Figure 2a, but not necessarily from the configuration in Figure 2b. Far from being competitive, the delivered price system in Figure 2a *derives from the exploitation of locational rents*!

A further peculiar result is readily evident with the BPP of equation (2). The low-cost producers just break even, but the high-cost nonbase-point firm that has a local monopoly position can earn supernormal profit or locational rents. Haddock's argument (p. 304) in this respect is in error. The locational rents arise because multiple-firm entry at the distant site is barred by a combination of demand and cost conditions. In particular, sufficient demand and revenues exist to generate returns for one firm that exceed opportunity costs, but entry of another firm would mean that both sellers incur economic losses. These conditions are exogeneous to the producer at the distant site. Such indivisibility has been discussed by Nicholas Kaldor (1935), Harold Demsetz (1964), and, more recently, B. Curtis Eaton and Richard G. Lipsey (1978).

One further concern of Haddock was to use BPP to present an efficiency argument for cross-hauling. But it is clear that cross-hauling does not require basing-point pricing. There is exactly the same form of cross-

[3] In the TV analysis, two competing producers will suffice.

hauling under the pricing equilibrium in equation (1), with firm *II* shipping to consumers to its left, and firm *I* shipping to consumers to the right of firm *II*. Simply put, differences in production costs and economies of long-haul transportation are sufficient to generate this kind of cross-hauling.[4]

In addition to the aforenoted theoretical criticisms, the idea of a competitive BPP system is further subject to the most basic criticism that it misinterprets the system. In particular, the BPP equilibrium under competitive conditions is justified on the grounds that it will allow the firm at site *II* to survive when f.o.b. pricing would not.[5] This result requires that firm *II* supply all demand in its natural market area: defined in Figure 1 as the area in which $m_{II}(x) < m_I(x)$ —region *ee'*. With such an assumption, BPP always generates greater total revenue than will f.o.b. pricing.[6] There will, therefore, exist some range of cost and demand conditions such that BPP will allow firm *II* to at least break even, while f.o.b. pricing will not. But this is not in the spirit of the BPP system. As Machlup (1949) states: "[the] basing-point technique of pricing makes it possible for any number of sellers, no matter where they are located and without any communication with each other, to quote identical delivered prices for any quantity of the product in standardized qualities and specifications" (p. 7).

Assume then that BPP leads to firm *II* sharing its market with the firms at site *I*. We know there must be at least two producers at site *I* for the subject BPP system to emerge. It follows that firm *II* can at the most expect only one-third of the total sales in the region *ee'* under the BPP system. By contrast, f.o.b. pricing will secure all demand for firm *II* in a region somewhat smaller than *ee'*. As Melvin L. Greenhut (1956, pp. 313, 314; 1970, Chapter 7) points out, once such market sharing is considered, f.o.b. pricing can dominate basing-point pricing and would if costs corresponded to those of Figure 1. Indeed, the classical form of spatial price discrimination by the distant firm, where its delivered price increases in *each direction* from its site, can dominate the f.o.b. pricing alternative; moreover, it would not appear to be predatory, as is the case for the systematic lowering of delivered prices in region *ee'* at some market points located at greater distances from the seller than others.

III. Conclusions

In order for a competitive simple basing-point price system to hold, it is necessary that a single firm located at a distance from a competitive production center exploit its local market power, collect locational rents, and alone sell over its dominated market space. This composite is a necessary condition; without its fulfillment, the *noncooperative* base-point price system would not exist. It follows in light of the requisite local market power at a distance that BPP is a noncompetitive system. Moreover, when such power exists, other delivered pricing forms will be used in the absence of collusion.

This paper thus contends that traditional views on the subject are more appropriate. (See Greenhut, 1956, Appendix V.) When base-point pricing is followed, it is attributable to (i) conscious parallelism of action, where some firms feel obliged (compelled) to adhere to the price schedule of others or (ii) stems from an outright collusive conspiracy whose design is to maximize profits *subject to* cartel policing and enforcement costs. Such costs are an intrinsic part of the system since a cartel must be organized and its agreement enforced in the presence of incentives to cheat. As many have pointed out (for example, Hoover, 1937, p. 190; Machlup, 1949, pp. 131–136; George

[4] Actually, even more extensive cross-hauling involving simultaneous supply of an area by two or more firms is easily explained with spatial competition models involving price discrimination (for example, John Greenhut and Melvin L. Greenhut, 1975). Base-point pricing is not a necessary condition for cross-hauling, or for freight absorption, so Haddock is clearly correct in emphasizing that these phenomena are not evidence of collusion. But significantly, they do not necessarily imply base-point pricing either.

[5] Actually, of course, the choice is more likely to involve discriminatory pricing than f.o.b. pricing.

[6] With trivial exception of some peculiar demand functions.

J. Stigler, 1979, pp. 1147–1148; Frederic M. Scherer, 1980, pp. 329–330), a base-point system is much less costly to police and enforce than are other spatial pricing arrangements. This helps explain its use instead of other organized delivered price systems that would be more profitable in the absence of policing and enforcement costs.

The fact that base-point pricing generates locational rents for distant firms also indicates why this system might characterize a cartel notwithstanding the existence of more profitable arrangements, such as f.o.b. pricing or spatial discriminatory pricing. These other more profitable systems may well induce competitive entry. It follows that if base-point pricing deters entry and in the process protects existing rents, it may actually be the most desirable arrangement that the cartel can adopt.

REFERENCES

Demsetz, Harold, "The Welfare Implications of Monopolistic Competition," *Economic Journal*, September 1964, *74*, 623–41.

Eaton, B. Curtis and Lipsey, Richard G., "Freedom of Entry and Existence of Pure Profit," *Economic Journal*, September 1978, *88*, 455–69.

Greenhut, John and Greenhut, Melvin L., "Spatial Price Discrimination, Competition and Locational Effects," *Economica*, November 1975, *42*, 401–19.

Greenhut, Melvin L., *Plant Location in Theory and in Practice*, Chapel Hill: University of North Carolina Press, 1956; 4th printing, Westport, CT: Greenwood Press, 1982.

______, *A Theory of the Firm in Economic Space*. New York: Appleton Century, 1970; 2nd printing, Austin, TX: Lone Star Publishing, 1974.

Haddock, David D., "Basing-Point Pricing; Competitive vs. Collusive Theories," *American Economic Review*, June 1982, *72*, 289–306.

Hoover, Edgar M., "Spatial Price Discrimination," *Review of Economic Studies*, 1937, *4*, 182–91.

Kaldor, Nicholas, "Market Imperfections and Excess Capacity," *Economica*, February 1935, *2*, 35–50.

Machlup, Fritz, *The Basing Point System: An Economic Analysis of a Controversial Pricing Practice*, Philadelphia: Blakiston, 1949.

______, *The Political Economy of Monopoly: Business, Labor and Government Policies*, Baltimore: Johns Hopkins University Press, 1952, esp. chs. 4 and 5.

Scherer, Frederic M., *Industrial Market Structure and Economic Performance*, Chicago: Rand McNally, 1980.

Stigler, George J., "A Theory of Delivered Price Systems," *American Economic Review*, December 1949, *39*, 1143–1159.

Thisse, Jacques F. and Vives, Xavier, "On the Strategic Choice of Spatial Price Policy," *American Economic Review*, March 1988, *78*, 122–37.

Wilcox, Clair, *Public Policies Toward Business*, Homewood, IL: Richard D. Irwin, 1955.

Government Debt, Government Spending, and Private Sector Behavior Revisited: Comment

By Martin Feldstein and Douglas W. Elmendorf*

Perhaps no issue has generated as much controversy among economists in the past decade as the proposition that an increase in the government deficit induces an equal offsetting increase in private saving. The truth of this so-called Ricardian equivalence proposition is central to whether budget deficits reduce capital accumulation, to the feasibility of expansionary tax reductions, and to the effects of Social Security on private saving and aggregate capital accumulation.

Although the basic idea that the future tax liabilities associated with government deficits and debt induce individuals to increase their saving has been around since the time of David Ricardo and was treated explicitly by Don Patinkin (1965), Martin Bailey (1971) and Levis Kochin (1974), the current debate was launched by Robert Barro (1974). The voluminous theoretical literature of recent years has shown that complete Ricardian equivalence would be expected to prevail only under very special conditions; see Douglas Bernheim (1987) for an especially useful survey and analysis. But the theoretical restrictiveness of the assumptions required for complete Ricardian equivalence does not constitute a practical refutation. Defenders of Ricardian equivalence can argue that the theory is only an approximation and can claim that, although the stringent conditions required for complete Ricardian equivalence do not hold, the economy's behavior in practice is close to the predictions of Ricardian equivalence.

There are two key empirical questions. The first is whether a higher level of taxes (with government spending constant) induces individuals to reduce their spending on consumption as traditional theory holds or has no effect on consumer spending as the Ricardian equivalence proposition predicts. The second deals with the effect of government outlays on goods and services. Although the absence of a negative effect on consumer spending of such government outlays is clearly contrary to the Ricardian equivalence proposition, the existence of a moderate negative effect of government outlays on consumer spending is not in itself evidence in favor of the Ricardian equivalence proposition that individuals increase their saving to finance anticipated debt service. As Feldstein (1982) explained, consumers may correctly believe that a rise in current government spending is a good indicator of a higher level of future government spending. Once a program is launched or budgets increased, the process is unlikely to be reversed. An increase in current government spending is therefore a good indication that future taxes will have to be higher to finance a higher level of future government spending. Individuals may rationally reduce their own spending when government outlays increase without a concurrent increase in taxes because they anticipate higher future taxes to finance higher future government spending even if they give little or no weight to the debt service implications of the current deficit.

The strongest direct evidence in favor of Ricardian equivalence is Roger Kormendi's 1983 article in the *American Economic Review*. He presents consumption regression equations that relate an estimate of consumption[1] to net national product, wealth,

*Martin Feldstein is Professor of Economics at Harvard University and President of the National Bureau of Economic Research. Douglas Elmendorf is Assistant Professor of Economics at Harvard University. We are grateful to Greg Mankiw and Lawrence Summers for comments on an earlier draft. The research reported here is part of the NBER study of the Government Budget and the Private Economy.

[1] Kormendi defines consumption as the sum of current expenditures on services and nondurables plus 10

government debt, government spending on goods and services, taxes, transfers, corporate retained earnings, and government interest payments. His parameter estimates appear to show that an increase in taxes does not affect consumption at all while an increase in government spending on goods and services does reduce consumption. Thus whatever the theoretical shortcomings of the Ricardian equivalence theory, it would appear from Kormendi's results that in practice consumers behave as the Ricardian equivalence theory predicts.[2]

Kormendi's analysis has been criticized by James Barth et al. (1986) and by Franco Modigliani and Arlie Sterling (1986). Although we are unconvinced by Kormendi's analysis, we do not find that either of those comments is a persuasive refutation of the Kormendi study. While Barth et al. conclude that their results "raise sufficient questions about the robustness and interpretation of Kormendi's original findings that more empirical work in this important research area is clearly needed" (p. 1165), their estimates generally support Kormendi's principal finding that consumer spending is sensitive to government outlays on goods and services but not to taxes. Their analysis extends the Kormendi sample through 1983, separates federal government debt from state and local debt, estimates for alternative subperiods, and tries substituting the par value of government debt for the market value of the debt even though they recognize that the latter is the conceptually appropriate measure. The only estimates in which the effect of government spending is not at least marginally significantly negative are in equations estimated for the postwar period that contain the theoretically inappropriate par value of the government debt. In no equation do taxes have a significant negative effect.

The essential feature of the Modigliani and Sterling analysis is to replace the separate tax, transfer, and government interest variables with a combined "net tax" variable that is equal to government taxes net of transfers including government net domestic interest payments. The sum of the distributed lag coefficients of this net tax variable is significantly negative in the Modigliani-Sterling consumer expenditure equations and is not significantly different from the sum of the lag coefficients of net national product. In this specification the coefficient of the government spending variable is small and not significantly different from zero. Modigliani and Sterling interpret their estimates as "strikingly and unmistakenly consistent with a "Life Horizon"–Life Cycle approach to consumption behavior and equally inconsistent with the infinite horizon Ricardian Equivalence Proposition formulation" (p. 1178).

Unfortunately, however, Modigliani and Sterling do not provide an explicit test of the effect of taxes per se on consumption but only of the combined "net tax" variable. Since the coefficient of the transfers variable in the original Kormendi analysis was positive, large, and statistically significant, it is not surprising that the variable created by subtracting transfers from taxes has a coefficient that is negative, large and statistically significant. This should be expected regardless of any additional changes in variable definitions or estimation procedures. Although the variable is correctly labeled as "net taxes," its coefficient is essentially an indication of the effect of transfers. Similarly, although Modigliani and Sterling do present two specifications that include transfers as a separate variable, in both of those equations they constrain the coefficient of

percent of current expenditures on consumer durables and 30 percent of the stock of consumer durables.

[2] There have of course been other tests of Ricardian equivalence. John Seater and Roberto Mariano (1985) interpret their evidence as supporting Ricardian equivalence while the evidence of Michael Darby et al. (1987) and of Feldstein (1982) is inconsistent with Ricardian equivalence. There have also been indirect tests of Ricardian equivalence based on examining the effects of government deficits on real interest rates and on the exchange rate. The results of these tests have been mixed. Some researchers, including Charles Plosser (1982) and Paul Evans (1985), found that budget deficits do not change interest rates or the value of the dollar while others, including Feldstein (1986) and Michael Hutchison and Adrian Throop (1985), have found the opposite. A discussion of these articles lies beyond the scope of this paper.

the net tax variable to equal the coefficient of net national product, so no separate estimates of the effects of taxes and transfers can be inferred.

These objections do not detract from the force of the Modigliani-Sterling argument that transfers are negative taxes and that the two should therefore be treated symmetrically in any analysis of Ricardian equivalence. If this is accepted, Kormendi's own estimates of the effect of transfers on consumption provide a strong refutation of the Ricardian equivalence proposition. Kormendi's reply that transfers are received by a subgroup of the population that is liquidity constrained suggests at a minimum that not all taxes and taxpayers should satisfy the Ricardian equivalence proposition.

We are nevertheless left without any direct test of Kormendi's conclusion that government spending depresses consumer spending while taxes do not. There are two ways in which Kormendi's research could be subject to further analysis. The first would be to develop a more general model of which Kormendi's is a special case and to evaluate the relative importance of government spending and taxes in that more general model. Such a model might include the real net interest rate, alternative measures of household wealth, the age distribution of the population, the income distribution, and other variables that could influence consumer spending but that are not part of the Kormendi analysis. The alternative and more modest but direct approach followed in the present paper is to see whether Kormendi's results remain when his equation is reestimated by different statistical techniques, using different functional forms, and for periods that exclude the World War II years, when consumer spending was constrained by shortages and rationing. When this is done, the results are quite contrary to Kormendi's and reject the Ricardian equivalence proposition.

More specifically, the present paper shows that Kormendi's results are misleading and cannot be sustained when the war years 1941 through 1946 are excluded from the sample. Omitting this period of wartime shortages and rationing and estimating for the period through 1985 reverses Kormendi's principal finding and shows that higher taxes depress consumption while an increase in government spending has no significant effect on consumption. This conclusion is confirmed when Kormendi's procedure of estimating in first-difference form is replaced by a more appropriate estimation using a first-order autoregressive transformation of an equation specified in levels.[3]

The paper begins (Section I) by replicating Kormendi's estimates and extending the end of the sample period from 1976 to 1985. Section II then shows the critical importance of Kormendi's practice of including the war years. Section III examines the effect of excluding state and local governments and focusing on federal government spending, taxes, transfers, and debt. The fourth section respecifies the equation in ratio form; this reduces the problem of collinearity and produces coefficients that imply stronger tax effects (with smaller standard errors) and no effect of government spending. The fifth section looks at estimates for the postwar period only, while the sixth section presents instrumental variable estimates that treat the current values of taxes, transfers, government spending, and NNP as endogenous. There is a final concluding section.

I. Replication and Basic Variations

Table 1 presents alternative estimates of Kormendi's basic specification relating consumer spending to net national product (Y_t), lagged net national product (Y_{t-1}), government spending on goods and services (GS_t),

[3] A proponent of Ricardian equivalence might of course say that this reversal of Kormendi's empirical findings does not contradict Ricardian equivalence since current taxes may be a proxy for future government spending. This line of argument, carried to the extreme, would make it impossible to refute Ricardian equivalence with estimates of consumer behavior. There is also a conceptual problem since taxes finance not only government spending on goods and services but also the empirically more important transfer payments and debt service. Moreover, regression equations estimated with our data (and similar equations estimated by Modigliani and Sterling) fail to find any predictive effect of current taxes on future spending.

TABLE 1—REPLICATION AND BASIC VARIATIONS OF THE KORMENDI SPECIFICATION

Equation	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)	(1.6)	(1.7)	(1.8)
Sample	1931–76	1931–76	1931–76	1931–85	1931–40 1947–85	1931–40 1947–84	1931–40 1947–85	1931–85
Data*	1	2	3	3	3	2	3	3
Estimation	FD	FD	FD	FD	FD	FD	AR1	AR1
Y_t	0.29	0.30	0.33	0.30	0.31	0.30	0.31	0.30
	(0.04)	(0.05)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.07)
Y_{t-1}	0.07	0.11	0.08	0.09	0.09	0.09	0.09	0.11
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
GS_t	−0.23	−0.15	−0.12	−0.10	0.05	0.03	0.13	−0.11
	(0.02)	(0.03)	(0.03)	(0.03)	(0.09)	(0.09)	(0.09)	(0.03)
TX_t	0.07	−0.08	−0.17	−0.15	−0.17	−0.13	−0.16	−0.07
	(0.08)	(0.10)	(0.11)	(0.11)	(0.13)	(0.13)	(0.12)	(0.12)
GB_t	−0.055	−0.039	−0.014	−0.008	−0.0004	0.014	0.081	−0.010
	(0.018)	(0.029)	(0.030)	(0.026)	(0.030)	(0.029)	(0.032)	(0.030)
TR_t	0.83	0.70	0.45	0.66	0.66	0.76	0.84	0.83
	(0.15)	(0.21)	(0.23)	(0.21)	(0.20)	(0.20)	(0.14)	(0.19)
W_t	0.025	0.021	0.033	0.015	0.014	0.016	0.020	0.024
	(0.008)	(0.011)	(0.013)	(0.009)	(0.008)	(0.007)	(0.006)	(0.008)
RE_t	0.10	−0.04	−0.13	−0.03	0.02	0.05	−0.04	0.16
	(0.11)	(0.15)	(0.18)	(0.17)	(0.18)	(0.17)	(0.18)	(0.20)
$GINT_t$	1.15	2.15	1.71	1.13	0.88	0.66	0.89	1.31
	(0.91)	(1.15)	(1.16)	(0.66)	(0.63)	(0.57)	(0.50)	(0.66)
R^2	0.91	0.83	0.79	0.75	0.79	0.83	0.999	0.999
SER	17.5	22.3	56.4	61.4	57.4	24.7	48.0	62.2
ρ	–	–	–	–	–	–	0.76	0.86
DW Stat.	n.a.	1.5	1.6	1.7	1.7	1.6	1.8	1.7

Note: The dependent variable is consumption. A constant term was included in the estimation but is not reported. Standard errors are in parentheses. See text for definition of variables.

*Data: [1]Kormendi [2]Pre-Benchmark [3]Post-Benchmark

total tax revenue (TX_t), government debt at market value (GB_t), transfer payments (TR_t), private wealth excluding government debt (W_t), corporate retained earnings adjusted for capital consumption allowances and inventory valuation (RE_t), government interest payments to persons and businesses ($GINT_t$), and a constant term. Standard errors of the parameter estimates are shown in parentheses beneath each coefficient.

Kormendi's original article presents detailed definitions and sources for these statistics. Since our purpose is to assess the sensitivity of Kormendi's conclusion to the inclusion of the World War II years, we have tried to stay as close as possible to his specification and his definitions in constructing the regression variables for our analysis. As far as we can tell, there are only two small differences between our definitions and Kormendi's. First, we use the personal consumption price deflator for all variables while Kormendi uses different price deflators for different variables. Second, we use the Federal Reserve Flow of Funds *Balance Sheets for the U.S. Economy, 1946–85* as the source of data on "domestic net assets" to measure private nonhuman wealth while Kormendi rescales estimates of several types of wealth presented in the *Survey of Current Business*; the Federal Reserve data have the virtue of being directly comparable to the estimates by Raymond Goldsmith used by Kormendi and by us for the earlier years of the sample. Kormendi's series on the market value of the government debt is extended using the method developed in James Butkiewicz (1983). A complete listing of the data is available on request.

Equation (1.1) of Table 1 reproduces the parameter estimates reported in Table 5 of Kormendi. This is the only equation reported by Kormendi that includes both the tax variable and the government debt variable. Note that it is estimated in first-difference form for the entire sample from 1931 through 1976, including the years of World War II. In this specification an extra dollar of government spending on goods and services reduces consumer outlays by 23 cents. By contrast, an extra dollar of tax revenue has no statistically significant effect; its coefficient is actually positive but less than its standard error. The coefficient of the government debt variable also has an implausible negative sign but is actually statistically significant. Although this might have been interpreted as evidence that the equation is misspecified or misestimated, Kormendi just notes that the negative coefficient is clearly contrary to the traditional theory that implies that government debt is a form of wealth that should have a positive effect on consumption.

Equation (1.2) represents our attempt to replicate the original Kormendi estimation for his sample period of 1931–1976. Since Kormendi's equation was estimated with the data available before the major 1985 benchmark revision of the national income accounts, we used pre-benchmark revision data to estimate equation (1.2). Although the results are not identical to equation (1.1) (presumably because of earlier revisions in the national income and wealth statistics and perhaps because of inadvertent differences in the way that the data are constructed), the coefficients are quite similar to Kormendi's results in equation (1.1). In particular, the government spending coefficient is -0.15 with a standard error of only 0.03, implying a significant effect of government spending on consumption, while the tax coefficient is only -0.08 with a standard error of 0.10. The coefficient of the debt variable remains negative and larger than its standard error.

Equation (1.3) repeats this estimation for the same sample period but with the most recent available data (as of July 1987). The primary effect of this data revision is to increase the coefficient of the tax variable to -0.17 with a standard error of 0.11, implying that the hypothesis that the true tax coefficient is zero or positive can be rejected at a 7 percent significance level. The coefficient of the government debt variable is still negative but drops to less than half its standard error. The government spending variable is reduced in size but is still substantially greater than its standard error.

Extending the sample period through 1985 (equation (1.4)) leaves these key parameter estimates essentially unchanged. The coefficient of the government-spending variable is significantly negative while the coefficient of the tax variable, although absolutely larger than the government-spending coefficient, is only significant at the 10 percent level.

II. Excluding the War Years

A crucial feature of the Kormendi estimates is that his sample includes the war years, when consumption was reduced by widespread rationing, by shortages of consumer durables and other consumer goods, and by patriotic appeals to purchase saving bonds. The personal saving rate jumped from 4.0 percent in 1940 to 10.9 percent in 1941 and to more than 20 percent in each of the next three years before subsiding to 19.2 percent in 1945 and 8.6 percent in 1946. The war years were also a time in which government spending rose much more than tax revenue. Real government spending on goods and services in 1982 dollars jumped from $150 billion in 1940 to $484 billion in 1942 and $791 billion in 1944, a level that has never been reached again. By contrast, real tax revenue only rose from $126 billion to $264 billion. The evidence presented below shows that the strong correlation between wartime government spending and the high saving rate caused by shortages and rationing causes a spurious negative relation between government spending and personal consumption in the sample as a whole.

Because of the very unusual nature of the consumer goods markets during the World War II period and the intense patriotic appeals for increased saving, the World War II years should be excluded in any regression analysis of saving behavior. Although Kormendi does present some estimates that exclude the war years, those equations never contain both the tax and government debt variables and therefore do not provide an explicit test of the Ricardian equivalence proposition.[4]

Equation (1.5) shows that when the six war years 1941 through 1946 are omitted, the remaining 49 observations tell a very different story. In particular, the coefficient of the government spending variable becomes very small and only about half as large as its standard error. In contrast, the coefficient of the tax variable is -0.17 with a standard error of 0.13 that implies that the null hypothesis of a zero or positive effect can be rejected at the 10 percent level. It is clear that including the war years produces very misleading results.

As a check that the reason for the very different conclusions implied by equations (1.1) and (1.5) is due to omitting the war years and not to the recent benchmark data revision, we have reestimated equation (1.5) using the data available before the 1985 benchmark revision. The results, presented in equation (1.6), are quite similar to those of (1.5) and indicate that it is the exclusion of the war years rather than the data revision that is critical.

Kormendi explains that he estimates the equations in first-difference form to reduce the risk of the spurious results that Clive Granger and Paul Newbold (1974) have shown can occur when the equations are estimated in level form and there is substantial serial correlation of the residuals. If the variables in levels are nonstationary, the hypothesis tests will be at a level different from that intended. However, using first-difference estimation is less efficient than estimation in level form with an autoregressive transformation. Moreover, estimation by first-differencing has some further disadvantages. If variables are measured with error the use of first difference estimation increases the errors in variables' bias. If the response of consumers to an explanatory variable is not immediate, the use of first-difference estimation can cause a substantial underestimation of its true effect. The remaining equations of Table 1 therefore present estimates in level form after an AR1 transformation based on the estimated autocorrelation coefficient.

Equation (1.7) is estimated after an AR1 transformation with an autocorrelation coefficient of 0.76. The results again indicate that taxes have a negative effect while government spending has a positive coefficient. The coefficient of the government debt variable has the correct positive sign and is more than twice its standard error; if anything, the coefficient is implausibly large. Equation (1.8) shows the effect of including the war years with the AR1 estimation; once again government spending becomes significant while taxes and government debt are insignificant. It is clear that the choice between first-difference estimation and an autoregressive transformation does not affect the conclusion that the evidence in favor of Ricardian equivalence rests on including the six years of World War II and that when these years are excluded Ricardian equivalence is clearly rejected.

III. Total Government or Federal Government

Federal taxes and spending are very different from the taxes and spending of state and local governments. Individuals can in principle avoid a very large part of the state and local taxes that they pay by moving to a different jurisdiction where they would also forego the benefits that higher tax dollars purchase. Moreover, while approximately 75 percent of federal government spending on goods and services is for national defense, the goods and services spending of state and local governments is for education and other personal services of the local voters. State and local debt is also different in kind from federal debt since the value of such area-specific debt will tend to be reflected in local property values.

[4] Kormendi does present one equation in Table 4 without the war years and with both government spending and taxes but without government debt. We have followed Kormendi and reestimated this equation in first-difference form but, unlike Kormendi, we did not obtain a significant effect of government spending. Our estimated coefficient of government spending was 0.02 with a standard error of 0.08; by comparison, the tax variable had a coefficient of -0.18 with a standard error of 0.12. The difference in our finding may well reflect revisions in the national income statistics.

TABLE 2—TOTAL GOVERNMENT VERSUS FEDERAL GOVERNMENT ONLY SELECTED COEFFICIENTS*

Equation	Sample	Estimation	Govt	GS_t	TX_t	GB_t	R^2	SER	ρ	DWS
2.1	1931–40	First	All	0.05	−0.17	−0.0004	0.79	57.4	–	1.7
	1947–85	Difference		(0.09)	(0.13)	(0.030)				
2.2	1931–40	First	Fed	0.02	−0.32	0.018	0.79	57.4	–	1.8
	1947–85	Difference		(0.09)	(0.14)	(0.034)				
2.3	1931–40	AR1	All	0.13	−0.16	0.081	0.9995	62.2	0.76	1.8
	1947–85			(0.09)	(0.12)	(0.032)				
2.4	1931–40	AR1	Fed	0.13	−0.31	0.150	0.9995	48.3	0.84	2.0
	1947–85			(0.09)	(0.13)	(0.044)				
2.5	1931–85	First	All	−0.10	−0.15	−0.008	0.75	61.4	–	1.7
		Difference		(0.03)	(0.11)	(0.026)				
2.6	1931–85	First	Fed	−0.09	−0.30	0.015	0.75	60.4	–	1.8
		Difference		(0.03)	(0.12)	(0.028)				

*The estimated equations include all of the variables presented in Table 1; standard errors are shown in parentheses.

The equations presented in Table 2 compare the key fiscal coefficients based on equations using the spending, taxes, and debt of all governments with the corresponding coefficients based on equations using federal government spending, taxes, and debt. Although only the coefficients of the three key fiscal variables are shown, they are obtained from estimates of the full equations of the form reported in Table 1; the full set of coefficients is available on request. Separate estimates are reported for the first difference and AR1 estimates.

Equations (2.1) and (2.2), estimated in first-difference form, indicate that federal taxes have a more powerful and statistically more significant effect on consumption than the taxes of state and local governments. The other coefficients are similar with insignificant effects of government spending and government debt. In the AR1 estimates reported in equations (2.3) and (2.4) the principal difference is again that the tax coefficient is larger and statistically more significant in the federal specification than in the equation for all governments. This is even true when the war years are included (equations (2.5) and (2.6)); the coefficient of taxes in the federal equation is −0.30 (with a standard error of 0.12) and therefore about three times as large as the coefficient of the government-spending variable. Thus when attention is restricted to the federal fiscal variables, taxes are important even when the war years are included although government spending is important only when the war years are included.

IV. A Ratio Specification

One of the problems in making precise inferences about the coefficient values is the collinearity among net national product and the various fiscal variables. The equations in Table 3 present an alternative specification that reduces the problem of collinearity by dividing each of the variables by the current value of net national product. These equations are estimated only for the specification without the war years.

Comparing the coefficients of Tables 2 and 3 shows that the coefficient of the tax variable is much larger both absolutely and relative to its standard error, in the ratio specification than in the linear specification. The coefficients of the government spending variable generally remain positive and not statistically significant. The coefficient of the debt variable is positive, larger than its standard error, and generally of a plausible magnitude.

The ratio specification thus provides even stronger evidence against the Ricardian equivalence proposition than the linear equations of Tables 1 and 2. The estimated autocorrelation parameters also show that there is less autocorrelation in the ratio form than in the linear form.

TABLE 3—VARIABLES AS RATIOS TO NNP 1931–40, 1947–85; SELECTED COEFFICIENTS*

Equation	Estimation	Govt	GS_t	TX_t	GB_t	R^2	SER	ρ	DWS
3.1	First Difference	All	0.004 (0.11)	−0.51 (0.19)	0.045 (0.032)	0.91	0.010	–	2.3
3.2	First Difference	Fed	−0.02 (0.10)	−0.77 (0.16)	0.058 (0.032)	0.93	0.009	–	2.5
3.3	AR1	All	0.13 (0.09)	−0.69 (0.08)	0.039 (0.024)	0.996	0.007	0.49	1.9
3.4	AR1	Fed	0.16 (0.08)	−0.69 (0.07)	0.079 (0.015)	0.997	0.007	0.37	1.9

*The estimated equations include all of the variables presented in Table 1; standard errors are shown in parentheses.

TABLE 4—POSTWAR SAMPLE: 1951–85 SELECTED COEFFICIENTS*

Equation	Functional Form	Estimation	Govt	GS_t	TX_t	GB_t	R^2	SER	ρ	DWS
4.1	Level	AR1	All	0.02 (0.13)	−0.19 (0.17)	0.078 (0.043)	0.999	47.2	0.52	1.8
4.2	Level	AR1	Fed	0.02 (0.14)	−0.36 (0.17)	0.113 (0.051)	0.999	46.4	0.53	2.0
4.3	Ratio	AR1	All	0.06 (0.14)	−0.62 (0.15)	0.136 (0.042)	0.95	0.006	0.51	2.0
4.4	Ratio	AR1	Fed	0.08 (0.12)	−0.60 (0.16)	0.190 (0.034)	0.96	0.005	0.53	2.0

*The estimated equations include all of the variables presented in Table 1; standard errors are shown in parentheses.

V. The Postwar Period

Combining the prewar and postwar years provides a sample of 49 usable observations and substantial variation in government spending, taxes, and national debt. It is nevertheless interesting to look at a more recent period that avoids the special conditions associated with the depression, the war, and the immediate postwar years. Table 4 presents estimates based on the 35 years from 1951 through 1985. The four equations include the level and NNP-ratio specifications and are estimated for the federal government only as well as for all governments combined. The coefficients are estimated with an AR1 transformation; the estimated autocorrelation coefficients are all approximately 0.50.

In all of the estimates, the coefficient of the government spending variable is small, positive, and much less than its standard error. In contrast, the coefficient of the tax variable is negative and larger than its standard error. With the federal fiscal variables the tax coefficient is quite large and more than twice its standard error. The coefficients of the government debt variable are always positive and generally more than double their standard errors but also typically larger than theory would suggest.

The estimates based on postwar data are therefore strongly contrary to the predictions of the Ricardian equivalence hypothesis.

VI. Instrumental Variable Estimation

The parameter values reported by Kormendi were all estimated without any attempt to deal with the problem of the endogeneity of net national product and of the fiscal variables. That is also the approach that has been followed until this point in the present paper. It is easy to believe, however,

TABLE 5—INSTRUMENTAL VARIABLE ESTIMATES FOR FULL SAMPLE 1934–40, 1947–85 FOR FAIR'S METHOD; 1935–40, 1947–85 FOR FIRST DIFFERENCE

				Selected Coefficients*						
Equation	Functional Form	Estimation	Govt	GS_t	TX_t	GB_t	R^2	SER	ρ	DWS
5.1	Level	First Difference	All	0.22 (0.21)	0.05 (0.25)	0.059 (0.039)	0.56	66.0	–	2.1
5.2	Level	Fair's Method	All	0.26 (0.10)	−0.13 (0.18)	0.039 (0.028)	0.999	56.2	0.84	1.6
5.3	Ratio	First Difference	All	−0.11 (0.22)	−0.39 (0.35)	0.053 (0.043)	0.80	0.012	–	1.7
5.4	Ratio	Fair's Method	All	0.15 (0.09)	−0.36 (0.15)	0.080 (0.023)	0.99	0.007	0.98	2.1
5.5	Level	First Difference	Fed	0.17 (0.25)	0.10 (0.31)	0.053 (0.056)	0.34	80.9	–	2.2
5.6	Level	Fair's Method	Fed	0.19 (0.10)	−0.46 (0.16)	0.052 (0.033)	0.999	59.0	0.88	1.9
5.7	Ratio	First Difference	Fed	−0.29 (0.27)	−0.51 (0.42)	0.065 (0.047)	0.78	0.012	–	1.7
5.8	Ratio	Fair's Method	Fed	0.07 (0.07)	−0.70 (0.08)	0.084 (0.015)	0.99	0.007	0.52	2.2

*The estimated equations include all of the variables presented in Table 1; standard errors are shown in parentheses.

that the current values of NNP and the fiscal variables will be correlated with the error of the consumption equation. A surprisingly large level of consumer spending would probably raise NNP and taxes and might reduce transfers and countercyclical government spending. To the extent that this is true, the ordinary least squares estimates would be biased and inconsistent.

An instrumental variable estimation procedure can provide consistent and asymptotically unbiased estimates in this context. The practical problem is to find satisfactory instrumental variables that are uncorrelated with the current disturbance to consumer spending but highly correlated with the endogenous explanatory variables. Variables like population that satisfy the first criterion completely generally do poorly by the second criterion. In the present study we have used as instruments the past values of the endogenous variables lagged two, three, and four years; that is, NNP_{t-2}, NNP_{t-3}, NNP_{t-4}, GS_{t-2}, GS_{t-3}, GS_{t-4}, TX_{t-2}, etc. These variables are clearly correlated with the fundamental movements and short-term trends in the corresponding variables but will only be correlated with the disturbance in the consumption equation to the extent that those disturbances have a high degree of serial correlation. While the instruments are not perfect, the use of instrumental variable estimation provides a check on the general qualitative properties of the ordinary least squares estimates.

Table 5 presents instrumental variable estimation for the entire sample with the war years omitted. In order to obtain the lagged values needed as instrumental variables it was necessary to drop the first four years from the sample; the sample therefore begins with 1935. In addition to instrumental variable estimates in first-difference form, we have also used Fair's method to combine instrumental variable estimation and a consistently estimated first-order autoregressive correction.[5] Although the instrumental variable estimates inevitably appear less precise than ordinary least squares estimates, the implications of Table 5 are very similar to those of the previous ordinary least squares

[5] Since our computer program could not apply Fair's method to a sample with a gap in the data, we have applied Fair's method to estimate our equation for the entire period from 1935 through 1985 but with individual dummy variables for each of the six war years.

TABLE 6—INSTRUMENTAL VARIABLE ESTIMATES FOR POSTWAR SAMPLE: 1951–85

				Selected Coefficients*						
Equation	Functional Form	Estimation	Govt	GS_t	TX_t	GE_t	R^2	SER	ρ	DWS
6.1	Level	First Difference	All	−0.16 (0.21)	0.24 (0.30)	−0.008 (0.061)	0.53	67.6	–	1.3
6.2	Level	Fair's Method	All	0.07 (0.10)	−0.11 (0.18)	0.030 (0.039)	0.999	57.7	0.996	1.6
6.3	Ratio	First Difference	All	0.21 (0.18)	−0.76 (0.35)	0.102 (0.058)	0.82	0.008	–	2.2
6.4	Ratio	Fair's Method	All	0.11 (0.09)	−0.66 (0.15)	0.113 (0.037)	0.95	0.006	0.82	2.3
6.5	Level	First Difference	Fed	−0.17 (0.23)	0.20 (0.34)	−0.017 (0.079)	0.48	70.8	–	1.2
6.6	Level	Fair's Method	Fed	0.14 (0.09)	−0.33 (0.15)	0.091 (0.048)	0.999	56.1	0.997	1.9
6.7	Ratio	First Difference	Fed	0.22 (0.16)	−0.81 (0.30)	0.137 (0.064)	0.84	0.008	–	2.4
6.8	Ratio	Fair's Method	Fed	0.15 (0.08)	−0.60 (0.15)	0.164 (0.045)	0.95	0.006	0.86	2.4

*The estimated equations include all of the variables presented in Table 1; standard errors are shown in parentheses.

estimates. The coefficient of government spending is generally positive and insignificant while the coefficient of the tax variable is generally negative and larger than its standard error. The coefficient of the government debt variable is always positive, generally greater than its standard error and of a roughly appropriate size. The use of Fair's method to correct for autocorrelation is generally helpful in obtaining more precise and more stable coefficients. As we noted with the OLS estimates of Tables 2 and 3, the results are generally stronger for the federal government fiscal variables and for the ratio specification.

Table 6 presents instrumental variable estimates for the postwar sample. The pattern of coefficients is again incompatible with the Ricardian equivalence proposition: generally positive and insignificant coefficients on government spending, negative and generally significant coefficients on the tax variable (with particularly strong effects in the ratio specification), and positive effects of the government debt.

VII. Concluding Comment

Because of the restrictive assumptions required to establish the theory of Ricardian equivalence, its relevance in practice is essentially an empirical question. Roger Kormendi's paper appeared to provide strong empirical support for Ricardian equivalence by showing that increases in government spending on goods and services depress consumer spending while changes in tax receipts have no effect on consumer spending.

The present study shows that Kormendi's results are a misleading implication of the experience during World War II, when shortages, rationing, and patriotic appeals to self-restraint caused an abnormally high rate of saving at the same time that the government deficit-financed a uniquely massive increase in defense spending. When those years are excluded from the sample, Kormendi's results are reversed.

The estimates presented here show that in the equation specified by Kormendi, but with the years 1941 through 1946 excluded, increases in tax receipts have had a substantial negative effect on consumption while increases in government spending on goods and services have had essentially no effect on consumption. This evidence is exactly the opposite of the implications of Ricardian equivalence. This conclusion is robust with respect to a variety of modifications in the way that the basic equation is estimated: using an AR1 correction to deal with serial correlation; limiting the analysis to the federal government's fiscal variables; respecifying the variables as ratios to net national product to reduce collinearity; estimating for the most recent 35 years instead of for the period since 1931; and using an instrumental variable procedure to reduce the problem of endogeneity. In each of these specifications, the results indicate that taxes depress consumer spending while government outlays on goods and services have either a smaller or a totally insignificant effect.

The present study has been limited to an analysis within the specification used by Kormendi. A different or more general specification might lead to different conclusions. But the present study has purposely been restricted to the Kormendi formulation because of the importance that has been attributed to Kormendi's evidence. The only proper inference that can be drawn from the present study is that Kormendi's own conclusion is wrong and that, within his own specification, the evidence decisively contradicts the Ricardian equivalence proposition and supports the conventional view that higher taxes reduce consumption and that budget deficits caused by tax reductions therefore depress national saving.

REFERENCES

Bailey, Martin J., *National Income and the Price Level*, New York: McGraw-Hill, 1971.

Barro, Robert J., "Are Government Bonds Net Wealth?" *Journal of Political Economy*, November/December 1974, *82*, 1095–1117.

Barth, James R., Iden, George and Russek, Frank S., "Government Debt, Government Spending, and Private Sector Behavior: Comment," *American Economic Review*, December 1986, *76*, 1158–67.

Bernheim, B. Douglas, "Ricardian Equivalence: An Evaluation of Theory and Evidence," *NBER Macroeconomics Annual*, 1987, *2*, 263–315.

Butkiewicz, James L., "The Market Value of Outstanding Government Debt: Comment," *Journal of Monetary Economics*, May 1983, *11*, 373–80.

Darby, Michael R., Gillingham, Robert and Greenlees, John S., "The Impact of Government Deficits on Personal and National Saving Rates," U.S. Department of the Treasury, Office of the Assistant Secretary for Economic Policy, Research Paper No. 8702.

Evans, Paul, "Do Large Deficits Produce High Interest Rates?" *American Economic Review*, March 1985, *75*, 68–87.

Feldstein, Martin S., "Government Deficits and Aggregate Demand," *Journal of Monetary Economics*, January 1982, *9*, 1–20.

______, "The Budget Deficit and the Dollar," *NBER Macroeconomics Annual*, 1986, *1*, 355–92.

Granger, C. W. J. and Newbold, Paul, "Spurious Regressions in Econometrics," *Journal of Econometrics*, July 1974, *2*, 111–20.

Hutchison, Michael M. and Throop, Adrian W., "U.S. Budget Deficits and the Real Value of the Dollar," *Economic Review of the Federal Reserve Bank of San Francisco*, Fall 1985, 26–43.

Kochin, Levis A., "Are Future Taxes Anticipated by Consumers?" *Journal of Money, Credit, and Banking*, August 1974, *6*, 385–94.

Kormendi, Roger C., "Government Debt, Government Spending, and Private Sector Behavior," *American Economic Review*, December 1983, *73*, 994–1010.

Modigliani, Franco and Sterling, Arlie, "Government Debt, Government Spending, and Private Sector Behavior: Comment," *American Economic Review*, December 1986, *76*, 1168–79.

Patinkin, Don, *Money, Interest, and Prices*, New York: Harper & Row, 1965.

Plosser, Charles I, "Government Financing Decisions and Asset Returns," *Journal of Monetary Economics*, May 1982, *9*, 325–52.

Seater, John J. and Mariano, Roberto S., "New Tests of the Life Cycle and Tax Discounting Hypotheses," *Journal of Monetary Economics*, March 1985, *15*, 195–215.

Government Debt, Government Spending, and Private Sector Behavior: A Further Comment

By Franco Modigliani and Arlie G. Sterling*

The response of Kormendi and Meguire (1986) to our earlier comment on Kormendi (1983) raises several issues which we feel need further clarification. They make three points:

i) we impose restrictions on the way transfers and interest enter the consumption function and, when those restrictions are released, our results regarding the Ricardian Equivalence Proposition (REP) are reversed;
ii) differencing is to be theoretically and empirically preferred in estimating the consumption function and
iii) that pre-War data add further support to their conclusions.

A. *The Role of Temporary Taxes*

KM base their conclusions on results using regression specifications similar to ours in all but one crucial respect, namely failure to take into account the role of temporary taxes. This specification can in no way be justified, either on theoretical or empirical grounds. Failure to take temporary taxes into account biases the results against the LCH, by lumping together taxes that should have a large effect on consumption with others that should have a small effect, even under the LCH.[1]

We test for the effect of temporary taxes by adding a variable "TEMPTAX" which is equal to the temporary taxes in every quarter in which such taxes were levied, and zero elsewhere. Since a temporary tax should reduce consumption by less than a permanent one, we should expect the coefficient of temptax to be significantly positive, though less than the absolute value of the coefficient of T. (Note that the variable T includes the transitory component.)

Once temporary taxes are taken properly into account, their role is found to be highly significant and the value of their coefficient is consistent with the hypothesis outlined above. Furthermore, the REP is decisively rejected. The results shown in the table illustrate our points. We begin in columns 1 and 2 by duplicating, for reference, the specification used in our earlier comment. We include for generality a correction for autocorrelation, though the coefficient (rho) is insignificant.[2] As in our earlier comment all the flow variables take the form of an Almon polynomial distributed lags of first degree, including the current year and the four preceding ones (hence each variable uses two degrees of freedom). Definitions of the variables are given in the notes to the table.

The results regarding the REP are essentially identical to those obtained before: consumers ignore government spending in evaluating their disposable income and wealth. This conclusion holds whether or not one imposes the "consistency condition"—

*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, and Marsoft Incorporated, Cambridge, MA 02139.

[1] One may ask why we attribute such an important role to transitory taxes and not to the transitory component that undoubtedly exists in other variables such as income or government expenditure. The answer is that the temporary nature of certain tax measures—especially those of 1969—was explicitly stated by the government, was widely understood, and the amounts involved were very large. We have endeavored to allow for transitory components in other variables through the device of distributed lags. Equivalently we assume that the (expected) permanent values of the variables can be approximated by a weighted average of current and previous values, with weights determined by the regression equation.

[2] We follow Kormendi and Meguire and ignore simultaneity issues. We note, however, that Sterling (1985) and Feldstein and Elmendorf (1990) treat the issue explicitly and find that an instrumental variable approach does not alter the conclusions.

TABLE 1—THE CONSUMPTION FUNCTION IN LEVEL FORM, 1952–84

	(1)	(2)*	(3)	(4)*	(5)	(6)*
Const.	-0.43	-0.44	-0.04	-0.00	-0.49	-0.39
	(0.05)	(0.06)	(0.10)	(0.12)	(0.09)	(0.09)
A	0.022	0.021	0.018	0.019	0.007	0.013
	(0.004)	(0.005)	(0.004)	(0.004)	(0.006)	(0.006)
G	0.096	0.093	0.050	0.065	0.115	0.147
	(0.021)	(0.031)	(0.022)	(0.030)	(0.030)	(0.027)
Y	0.93	0.94	0.79	0.74	0.97	0.77
	(0.02)	(0.07)	(0.04)	(0.06)	(0.03)	(0.07)
TEMPTAX	0.63	0.69	1.26	1.33		
	(0.19)	(0.35)	(0.22)	(0.35)		
DEF	0.19		0.09		-0.62	
	(0.07)		(0.17)		(0.19)	
T		-1.10		-0.78		-0.14
		(0.15)		(0.22)		(0.15)
GS		0.13		0.12		-0.28
		(0.14)		(0.18)		(0.19)
TR			0.54	0.63	0.29	0.49
			(0.08)	(0.11)	(0.11)	(0.11)
RGINT			0.32	0.38	1.70	1.65
			(0.87)	(0.43)	(0.46)	(0.37)
rho	0.04	0.07	-0.52	-0.54	-0.10	-0.30
	(0.20)	(0.21)	(0.18)	(0.19)	(0.21)	(0.21)
Se	0.0123	0.0127	0.0078	0.0078	0.0114	0.0098
Like	103.46	103.93	121.58	123.01	107.49	114.13

*In the columns marked with an asterisk Y is defined as net national product. In the remaining columns Y is defined as net national product less net taxes, as defined in the notes to the table.

The consumption function is in level form and all flow variables are estimated using a first degree four quarter (plus contemporaneous term) Almon lag. Rho is the autocorrelation correction factor, calculated using TSP's maximum likelihood method. Se is the standard error of the equation. Like is the value of the log-likelihood function. In significance tests twice the difference between the values in the constrained and unconstrained specifications is distributed chi-squared with degrees of freedom equal to the number of restrictions.

Variable Definitions:

Y represents income (see the asterisk note to the table).

T represents net taxes—total taxes less transfers and real net interest payments.

TEMPTAX represents temporary taxes, from the Economic Report of the President of 1969 and 1975, as defined in our earlier comment.

GS is government spending on goods and services.

TR is gross transfer payments by the government sector to the private sector (excluding interest).

RGINT is real net interest payments on general government debt, calculated by deducting the product of actual inflation and outstanding debt from nominal interest payments.

DEF is government spending on goods and services minus net taxes.

A is net worth of households (including government debt).

G represents the sum of federal, state, and local net financial liabilities at book value.

that the sum of the coefficients on government spending and taxes be equal and opposite to the propensity to consume from income before taxes. (That constraint is imposed in column 1 and released in column 2, and the consistency condition is not rejected.)

B. *Interest and Transfers*

In addition to dropping temporary taxes, the KM specification adds transfers and interest to equations (1) and (2). These variables are already included in net taxes, *T*, but we can see no objection to testing whether they might have an effect larger or smaller than that of other components of net taxes. Note in particular that under the REP hypothesis all transfers should have the same effect on consumption as all taxes, namely *no effect at all.*

Thus our analysis proceeds by keeping temporary taxes in the specification but adding real net interest and transfers, as shown in columns (3) and (4).

Comparing column (1) with (3) and (2) with (4), it is seen that, with or without the consistency condition (which is again not rejected by the data at the 5 percent level), there is a very large increase in the likelihood ratio. This evidence is consistent with the KM hypothesis that these variables may have effects on consumption different from that of taxes (though the coefficient of interest is not significant). But the results in columns (3) and (4) in no way support the REP. In (3) the deficit is positive—though insignificant and in (4) government spending receives a *positive* coefficient (but again completely insignificant). Thus KM's claim, that our conclusions regarding REP are reversed once the constraints we imposed on interest and transfers are released, is contradicted by these results.

As a comparison of columns (3) and (4) with columns (5) and (6) makes clear, KM came to their incorrect conclusion because they overlooked the role of temporary taxes. If temporary taxes are not included, the deficit and tax coefficients estimates could appear to support REP. But that omission is clearly rejected by the data—temporary taxes have a t-ratio of 5.7 in (3) and 3.8 in (4).

One final comment is in order with regard to the role of transfers. According to (3) and (4), the effect of transfer payments would be extraordinary indeed, amounting, according to equation (3), to $0.79 + 0.09 + 0.54$, or 1.42 dollars of consumption per dollar of transfer, with a hefty level of significance. If taken at face value, this estimate constitutes one more impressive piece of evidence rejecting REP, according to which the effect of transfers should be zero.

But we must point out that this estimate is absurdly high, both in level and relative to the estimated propensity to consume. In addition, these variables also result in an equally improbably high estimate of the coefficient of temporary taxes. We are therefore inclined to regard equations (3) and (4) as a statistical freak and to put credence in our original specifications (1) and (2). But, whether one prefers one specification or the other, REP stands soundly rejected.

C. *Differencing*

Kormendi and Meguire report that the residuals of their specifications—estimated in level form—show evidence of autocorrelation. As shown by the row labeled "rho" we also find some evidence of autocorrelation with the Kormendi and Meguire specification of columns (3) and (4). Strikingly, however, the residuals are negatively serially correlated and the value of rho is -0.52 and -0.54, respectively. In all cases the data reject the hypothesis that rho could be constrainted to equal 1.0, the value implicit in the differenced specification. In columns (5) and (6) the point estimate of rho is more than eight standard deviations from unity.

Engle and Granger (1987) show that, for the level form maximum likelihood methods we have used to be appropriate, it is sufficient to show that consumption and the explanatory variables used in our regressions are cointegrated. All the equations we report pass their first test for cointegration: the Durbin-Watson statistic of the untransformed equation is positive and so large (for example, in the untransformed version of equation (2) it is 1.90) that cointegration cannot be rejected. These results are consistent with those of Engle and Granger in their analysis of per capita consumption of nondurables and disposable income (in the work cited in Section 6).

Given the strong rejection of the differenced specification it is no surprise that the coefficients reported by Kormendi and Meguire are different from those obtained using the level specification shown here.

D. *The Irrelevance of Tests Including the Prewar Data*

In the tests reported above we have focused exclusively on the postwar span from

1952 to 1984, on the grounds that this sample is quite large and that no useful information can be obtained by reaching back to the years 1931 to 1951. The reason is that the bulk of that period—well over half—consists of (i) the chaotic years of the great depression, and (ii) the war years and the immediately surrounding ones. The depression period is clearly most unsuited to learn something about normal consumer behavior. And, as for the war years, it is well known that saving then was exceptionally high. This is likely to reflect patriotic motivations as well as the unavailability of many commodities. At the same time, that period is also characterized by a huge deficit, the largest in American history.

This combination clearly has the effect of biasing the coefficient estimates in favor of REP. Therefore one might be justified in including at most a handful of years, say 1940–41 and 1948–51. However the data that are available for the postwar period do not exist for the earlier period. Some information can be pieced together for these earlier years but it is clearly less reliable and not directly comparable with the data for the postwar period. Thus, in order to add the prewar years, KM ended up by not using the more reliable data available for the postwar period. We conclude therefore that estimates based on the combined period 1931–1984 should be disregarded.

E. *Conclusions*

We conclude that there are no grounds for changing our original proposition that the postwar data provide no support for the hypothesis that consumers consider government spending or the deficit when making their consumption decisions. Kormendi and Meguire arrive at a contrary conclusion by overlooking the theoretically and empirically important effects of temporary taxes, using an inefficient differenced specification, an unrepresentative sample, and less reliable data.

REFERENCES

Engle, R. F. and Granger, C. W. J., "Co-Integration and Error Correction, Representation, Estimation, and Testing," *Econometrica*, 1987, *55*(2), 251–76.

Feldstein, Martin S. and Elmendorf, Douglas W., "Government Debt, Government Spending and Private Sector Behavior: Comment," *American Economic Review*, June 1990, *80*, 589–99.

Kormendi, Roger C., "Government Debt, Government Spending, and Private Sector Behavior," *American Economic Review*, December 1983, *73*, 994–1011.

______ **and Meguire, Philip**, "Government Debt, Government Spending, and Private Sector Behavior: Reply," *American Economic Review*, December 1986, *73*, 1180–87.

Modigliani, Franco and Sterling, Arlie, "Government Debt, Government Spending, and Private Sector Behavior: Comment," *American Economic Review*, December 1986, *76*, 1168–79.

Sterling, Arlie G., "Public Debt and Private Wealth: The Role of Anticipated Taxes," unpublished doctoral thesis, Massachusetts Institute of Technology, April 1985.

Government Debt, Government Spending, and Private Sector Behavior: Reply and Update

By ROGER C. KORMENDI AND PHILIP MEGUIRE*

The preceding articles by Martin Feldstein and Douglas Elmendorf and by Franco Modigliani and Arlie Sterling give us a welcome opportunity to return to the effects of fiscal policy on private consumption. At stake in this debate, we believe, is a potential paradigm change—from what Kormendi (1983) termed the Standard Approach, which bases private consumption on disposable personal income, to what he termed the Consolidated Approach, which bases consumption on aggregate income, government spending, and transfer payments, each with separate effects. The Standard Approach excludes Ricardian equivalence a priori. The Consolidated Approach not only incorporates Ricardian equivalence but, in its augmented form, allows one to nest the various hypotheses associated with the two approaches.

Feldstein and Elmendorf argue that Kormendi's (1983) results (and implicitly those of Kormendi and Meguire, 1986), which reject the Standard Approach in favor of the Consolidated/Ricardian alternative, are not robust to the exclusion of data from World War II and other specification changes. Modigliani and Sterling argue that accounting for "temporary taxes" reverses our rebuttal of their 1986 comment. We first take up the challenge of Feldstein and Elmendorf before turning to Modigliani and Sterling. We then assess the validity of our preference for estimating in differences, by testing whether consumption, income, and the fiscal variables are cointegrated. Finally, we summarize what can be learned from the debate.

I. Feldstein and Elmendorf

Feldstein and Elmendorf focus on "two key empirical questions" in Kormendi (1983), namely, the relative effects of taxation and government spending on private consumption. Under Ricardian equivalence, (a) taxation should have no effect on private consumption, and (b) government spending, however, should have a negative effect, since spending summarizes the true resource burden of the government sector on the private sector. Under what Feldstein and Elmendorf term "traditional theory," taxation should have a negative effect and government spending should have no effect. These dichotomous predictions allow one to distinguish between Ricardian and traditional theories. Testing these predictions was central to Kormendi (1983) and Kormendi and Meguire (1986), whose results were strongly consistent with the Consolidated Approach in general, and with the above Ricardian predictions in particular.[1]

Feldstein and Elmendorf contend that Kormendi's (1983) results are not robust when data for World War II are omitted from the sample, while their results—strong negative effects of taxation on private consumption but no effects on government spending—are robust to a number of alter-

*The University of Michigan, School of Business Administration, Ann Arbor, MI 48109-1234, and Mid America Institute for Public Policy Research, Chicago, IL 60615.

[1] By allowing transfer payments and government spending to have consumption effects that depend upon their composition, the Consolidated Approach of Kormendi (1983) is more general than Ricardian equivalence. While a "strong" form of Ricardian equivalence implies that transfers should not affect the time path of aggregate consumption, Kormendi and Meguire (1986, p. 1186) show how the Consolidated Approach can reconcile the Ricardian and Life Cycle hypotheses via a suitable understanding of the role of transfers. See also Kormendi and Meguire (1989), who explore this reconciliation using a functional decomposition of transfers.

native specifications. In this section, we show that the results of Kormendi (1983) and Kormendi and Meguire (1986) are in fact fully robust to all of the alternative specifications proposed by Feldstein and Elmendorf. We then show how to reconcile our results with their apparent challenge.

A. *Two Theoretical Issues*

Feldstein and Elmendorf make two theoretical assertions that we wish to address before turning to their empirical work. First, they state that "...the absence of a negative effect of government outlays on consumer spending is clearly contrary to the Ricardian equivalence proposition." While this may be true under a strong version of Ricardian theory, in which government spending is implicitly either dissipative or utility-separable, it would not necessarily be true under the more general Consolidated Approach of Kormendi (1983). In particular, if the bulk of government spending were devoted to nondissipative investment projects that substituted for private investment but not for private consumption, then the Consolidated Approach would predict a zero effect of such spending on private consumption. Alternatively, if government spending were on goods and services that were at least partially complementary to (instead of substitutable for) private consumption, then current period government spending could have zero or positive effects on private consumption.

Second, Feldstein and Elmendorf assert that the existence of a "...moderate negative effect of government outlays on consumer spending is not in itself evidence in favor of Ricardian equivalence," but is possible under "traditional theory" as modified by Feldstein (1982). We quote Feldstein and Elmendorf:

> ...consumers may correctly believe that a rise in current government spending is a good indicator of a higher level of future government spending. Once a program is launched or budgets increased, the process is unlikely to be reversed. *An increase in current government spending is therefore a good indication that future taxes will have to be higher to finance a higher level of future government spending*. Individuals may rationally reduce their own spending when government outlays increase without a concurrent increase in taxes *because they anticipate higher future taxes to finance higher future government spending*. [emphasis added]

Ironically, this quotation embodies a line of reasoning virtually identical to that underlying the Ricardian equivalence proposition, namely, that the private sector correctly accounts for the future taxes implied by current fiscal policy.

B. *A Brief Review of Feldstein-Elmendorf's Empirical Results*

Feldstein and Elmendorf present in their Table 1 regressions that attempt first to replicate, and then to rebut, Kormendi's (1983) results while staying "...as close as possible to his specifications and definitions in constructing the regression variables." While they claim that their equation (1.2) replicates reasonably well Kormendi's (1983) equation (5.1), we point out that Barth, Iden, and Russek (1986, Table 1) achieve a much closer replication. Feldstein and Elmendorf then update their data base, incorporating the 1985 benchmark revisions of the national income accounts. Their equations (1.3) and (1.4) yield marginally significant negative coefficients for taxes while reducing the significant negative coefficient for government spending. In (1.5), which omits the years 1941–46, the negative effect of government spending vanishes, although the tax effect of -0.17 has a t-statistic of only -1.3.

In their Tables 1 through 6, Feldstein and Elmendorf explore the consequences of various combinations of four main specification changes: (1) estimating in levels with an AR1 correction, (2) using fiscal variables measured only over the federal government, (3) specifying all variables as ratios to NNP, and (4) estimating via instrumental variables. They also explore the effects of restricting the sample to the period 1951–85. The apparent force of Feldstein and Elmendorf's argument emerges only when two or

TABLE 1—UPDATED KORMENDI SPECIFICATION, WITH ALTERNATIVES SUGGESTED BY FELDSTEIN AND ELMENDORF

Equation	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)	(1.6)	(1.7)
Sample Period	1931–85	1931–40 1947–85	1930–40 1947–85	1931–40 1947–85	1931–40 1947–85	1935–40 1947–85	1951–85
Alternative Specification		Omit War	Level Data with AR1 Correction[a]	Federal Data Only	Ratio Form with $1/Y_t$	Instrumental Variables[c]	Postwar Data
Y_t	0.28	0.30	0.29	0.32	–[b]	0.22	0.31
	(0.06)	(0.06)	(0.06)	(0.06)		(0.10)	(0.11)
Y_{t-1}	0.07	0.09	0.09	0.07	0.05	0.06	0.12
	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.05)	(0.05)
GS	−0.26	−0.24	−0.24	−0.24	−0.25	−0.25	−0.22
	(0.03)	(0.07)	(0.08)	(0.07)	(0.06)	(0.11)	(0.11)
TX	0.06	0.05	0.14	−0.03	−0.12	0.17	0.00
	(0.10)	(0.12)	(0.12)	(0.13)	(0.12)	(0.22)	(0.16)
TR	0.73	0.85	1.10	0.78	0.68	0.67	0.81
	(0.21)	(0.21)	(0.17)	(0.24)	(0.14)	(0.40)	(0.29)
RGINT	0.03	0.31	0.37	0.39	0.09	0.29	0.44
	(0.07)	(0.10)	(0.09)	(0.12)	(0.06)	(0.19)	(0.30)
GB	−0.035	0.034	0.058	0.074	−0.021	0.059	0.052
	(0.024)	(0.032)	(0.033)	(0.034)	(0.025)	(0.043)	(0.045)
W	0.012	0.013	0.018	0.016	0.017	0.017	0.006
	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)	(0.010)	(0.010)
RE	0.17	0.07	0.09	0.00	−0.06	0.07	0.06
	(0.17)	(0.17)	(0.17)	(0.16)	(0.11)	(0.23)	(0.24)
R^2	0.766	0.815	0.976	0.822	0.967	0.680	0.706
SER	0.0285	0.0264	0.0260	0.0259	0.0057	0.0278	0.0266
DW	1.4	1.5	1.6	1.4	1.8	1.5	1.7

Note: All variables, including the dependent variable consumption, are in real per capita terms. We derive consumption, *Y*, and *GS* directly from real data; we obtain the other variables by dividing nominal data by the *NNP* deflator. *RGINT* is real government interest payments. All data are *post* the 1985 benchmark revision of the *NIPA*. For additional details on the data and variables, see the Data Appendix here and in Barth et al. (1986).

Standard errors of estimated coefficients are shown in parentheses. All regressions estimated with a constant (not reported). R^2 is *not* adjusted for degrees of freedom. *SER* = standard error of the regression; *DW* = Durbin-Watson statistic. Except for (1.3), all regressions were estimated over differenced data. In (1.4), we compute *GS*, *TX*, *TR*, *RGINT*, and *GB* from data for the federal government alone, ignoring grants-in-aid to state and local governments. In (1.5), we express all variables as ratios to *NNP*, with $1/Y_t$ included among the regressors.

[a] Estimated (Cochran-Orcutt) residual first-order serial correlation coefficient in 0.88.

[b] Coefficient of $1/Y_t$ is 0.77 (0.12). The constant is 0.0025 (0.0012).

[c] Instrument list: constant, *W*, *GB*, *RGINT*, *RE*, and lags 2 to 4 of *Y*, *GS*, *TX*, and *TR*.

more of these alternative specifications are combined. Whereas the tax coefficient is negative but only marginally significant in their Table 1, that coefficient is often large and negative in Tables 2 through 6.

Taken as a whole, Tables 1 through 6 in Feldstein and Elmendorf seem to contain strongly anti-Ricardian estimates. If these results were to stand up to detailed scrutiny, we would gladly yield to the weight of the evidence.

C. *The Robustness of the Consolidated Approach to the Feldstein-Elmendorf Specification Changes*

Table 1 presents the Feldstein and Elmendorf alternative specifications estimated using our own data. In particular, we follow Barth, Iden, and Russek (1986) and Kormendi and Meguire (1986) in our variable definitions, and apply these to the revised national income accounts. Further details on the data and the construction of the regression variables are given in the Data Appendix to this paper.[2]

Equation (1.1) in Table 1 reestimates the canonical Augmented Consolidated specification of Table 4, column 2, in Kormendi and Meguire (1986), over the 1931–1985 period. The results are very close to those we published in 1986; if anything, they provide even stronger evidence in favor of Ricardian equivalence and the Consolidated Approach. Note especially that the spending and tax coefficients in our (1.1) are -0.26 and 0.06, respectively, whereas these coefficients in Feldstein and Elmendorf's (1.4) are -0.10 and -0.15.

In equations (1.2) through (1.6), we delete the World War II years, first for the canonical specification (1.2), and then for Feldstein and Elmendorf's four main alternative specifications taken one at a time (1.3–1.6). Equation (1.7) is (1.1) estimated over the 1951–85 period. These estimates all bear out the Consolidated Approach, that is, large, negative and significant effects of government spending combined with small and insignificant effects of taxes.[3]

D. *Reconciling Feldstein-Elmendorf's Results with Those of Table 1*

The results in Table 1 above are both dramatically different from those reported by Feldstein and Elmendorf and fully consistent with our earlier results. In Table 2, we show the steps that bridge the gap between their results and ours. We report the government spending and tax coefficients[4] (along with their standard errors) for the five key specifications, each estimated with and without the war years: the "basic" differenced specification of Kormendi (1983), level data with an AR1 correction, federal fiscal variables only, ratio-to-NNP form, and instrumental variables. Under each specification, we show the number of the corresponding Feldstein-Elmendorf regression.

Column 1 of Table 2 presents our attempt to replicate Feldstein and Elmendorf. With the exception of the instrumental variables (IV) estimates, our replications are very close (and our IV replications are generally more favorable to their hypotheses than their own IV results are). Taken as a whole, however, the results in column 1 show that, contrary to Feldstein-Elmendorf's claims, there is in fact little material difference between estimates for the full period and those with World War II omitted. Both samples consistently yield negative coefficients for the tax variable. Although the government spending coefficients are significantly negative in the full sample, they are nearly always smaller in magnitude than the tax coefficients. Moreover, as discussed above, Feldstein and Elmendorf themselves assert that negative government spending coefficients are not necessarily evidence in favor of Ricardian equivalence.

Hence Feldstein and Elmendorf's anti-Ricardian results are not the result of omitting the war years from the sample.[5] Rather, their results stem from two seemingly inconsequential choices in the construction of their data.

We now address these choices in turn. First, Feldstein and Elmendorf's deflate nominal data by the consumption expenditures deflator. In making this choice, they overlooked the availability of published real

[2] See also the meticulous Data Appendix of Barth, Iden, and Russek (1986).

[3] In specifying their ratio-to-NNP specification (1.5), Feldstein and Elmendorf do not include $1/Y_t$, and hence fail to take the constant into account. The t-statistic of $1/Y_t$ is 6.5, and hence we include it in all the ratio form results reported here. In only a few cases, however, does failing to include $1/Y_t$ in (1.5) have much effect on the coefficients of either government spending or taxes.

[4] In the regressions underlying Table 2, the coefficients for government debt were generally insignificantly different from zero, with as many negative estimates as positive ones.

[5] To those who have read Kormendi (1983), Kormendi and Meguire (1986), or Barth, Iden, and Russek (1986) carefully, it should come as no surprise that the war years are not the key factor driving Feldstein and Elmendorf's results. These three papers investigated in detail the sensitivity of estimates to various subperiods, including ones that did not include the war years, and found no material evidence of inhomogeneity, especially for the government spending and tax coefficients.

TABLE 2—RECONCILING FELDSTEIN AND ELMENDORF'S *GS* AND *TX* COEFFICIENTS WITH TABLE 1

	(1)		(2)		(3)		(4)		(5)	
Author:	FE								KM	
Improved Variable Definitions	No		No		Yes		Yes		Yes	
Use Published Real Data	No		Yes		No		Yes		Yes	
Deflator for Nominal Data	Consumption Expenditures		Consumption Expenditures		Consumption		Consumption		*NNP*	
Specification (Number of Corresponding Estimate in FE)	*GS*	*TX*	*GS*	*TX*	*GS*	*TX*	*GS*	*TX*	*GS*	*TX*
1931–85										
Basic	−0.10	−0.14	−0.21	0.01	−0.22	0.02	−0.26	0.15	−0.26	0.07
(1.4)	(0.03)	(0.11)	(0.03)	(0.10)	(0.03)	(0.10)	(0.03)	(0.10)	(0.03)	(0.11)
AR1	−0.11	−0.09	−0.24[a]	0.16[a]	−0.23	0.06	−0.28	0.27	−0.30	0.16
(1.8)	(0.03)	(0.12)	(0.03)	(0.11)	(0.03)	(0.11)	(0.03)	(0.10)	(0.03)	(0.12)
Federal only	−0.09	−0.28	−0.21	−0.10	−0.20	−0.04	−0.25	0.08	−0.25	−0.05
(2.6)	(0.03)	(0.11)	(0.03)	(0.11)	(0.03)	(0.12)	(0.03)	(0.13)	(0.03)	(0.11)
Ratio[b]	−0.10	−0.35	−0.23	−0.15	−0.20	−0.12	−0.27	0.01	−0.26	−0.10
(*NR*)	(0.03)	(0.11)	(0.02)	(0.09)	(0.02)	(0.10)	(0.02)	(0.10)	(0.02)	(0.09)
Instrumental	−0.07	−0.18	−0.22	−0.01	−0.26	−0.04	−0.28	0.13	−0.26	0.01
Variables (*NP*)	(0.06)	(0.30)	(0.06)	(0.25)	(0.08)	(0.28)	(0.08)	(0.30)	(0.06)	(0.27)
1931–40, 1947–85										
Basic	0.05	−0.16	−0.17	0.04	−0.08	−0.03	−0.23	0.10	−0.23	0.06
(1.5)	(0.09)	(0.13)	(0.08)	(0.13)	(0.09)	(0.12)	(0.09)	(0.12)	(0.08)	(0.12)
AR1	0.11	−0.13	−0.16	0.15	−0.02	−0.01	−0.19	0.20	−0.23	0.16
(1.7)	(0.08)	(0.12)	(0.08)	(0.11)	(0.08)	(0.11)	(0.09)	(0.11)	(0.08)	(0.12)
Federal only	0.03	−0.32	−0.19	−0.09	−0.10	−0.07	−0.24	0.01	−0.23	−0.03
(2.2)	(0.09)	(0.14)	(0.08)	(0.15)	(0.08)	(0.14)	(0.08)	(0.15)	(0.07)	(0.14)
Ratio[b]	−0.05	−0.47	−0.24	−0.19	−0.08	−0.22	−0.26	−0.08	−0.24	−0.17
(*NR*)	(0.08)	(0.13)	(0.06)	(0.12)	(0.07)	(0.13)	(0.06)	(0.13)	(0.06)	(0.12)
Instrumental	0.01	−0.03	−0.24	0.06	−0.13	0.08	−0.24	0.22	−0.26	0.17
Variables (5.1)	(0.16)	(0.22)	(0.18)	(0.24)	(0.15)	(0.24)	(0.18)	(0.27)	(0.12)	(0.23)

Note: Standard errors of estimated coefficients are shown in parentheses. We will make available upon request the complete regression corresponding to any entry in this table. FE = Feldstein and Elmendorf (1990); KM = Kormendi and Meguire (1986); *NR* = not reported by FE.

In (1), we used FE's actual values (courtesy of Douglas Elmendorf) for consumption (the dependent variable), population, *GB* and *W*. We use their values for *W* throughout this table, so that the results under (5) do not correspond exactly to those in Table 1. For further details on the data and computation of the variables, see the Data Appendix and the note to Table 1.

Except for the data differences described in the preceding paragraph, the Basic specification is the same as (1.1) and (1.2) in Table 1. We compute AR1 estimates from levels data with a correction for residual serial correlation, either maximum likelihood (1931–85) or Cochrane-Orcutt (otherwise). All other specifications are estimated using OLS over first differences and are described in the note to Table 1.

[a] Maximum likelihood algorithm failed to converge.

[b] With $1/Y_t$.

(constant dollar) data for net national product, government spending, and the components of private consumption, assuming explicitly that doing so involved only "small differences." Moreover, since the dependent variable is not a measure of consumption expenditures but rather of consumption flow, the implicit deflator for the latter would be the more consistent choice. Whether to use the available real data or data deflated by some other deflator is an instance of the classical "index number problem," and as such cannot be resolved apriori.[6] However, in light of Feldstein and Elmendorf's stated purpose "...to assess the sensitivity of Kormendi's conclusion to the inclusion of the World War II years, [while attempting] *to stay as close as possible to his specification and his definitions*" (emphasis added), we now consider the sensitivity of their results to deviating from our earlier work in this respect.

To this end, we present in column 2 of Table 2 estimates identical to those in column 1, except that Y, GS, and consumption are constructed from published real data. The results for all specifications are now fully consistent with the Consolidated Approach, whether or not the World War II years are excluded. Government spending has significant negative effects of about the magnitude uncovered by Kormendi (1983), Barth, Iden, and Russek (1986), and Kormendi and Meguire (1986), while taxes have small and insignificant effects of mixed sign.

Second, Feldstein and Elmendorf do not implement the improved data and variable definitions that emerged from the 1986 debate. For example, Barth, Iden, and Russek (1986) meticulously reconstructed, corrected, and extended Kormendi's (1983) original data and variable definitions. Further, Modigliani and Sterling (1986) argued correctly that the Fisherian inflation premium should be netted from nominal interest payments to obtain a measure of real government interest payments, *RGINT*. In Kormendi and Meguire (1986), we acknowledged these improvements and added some of our own, including incorporating Cox's (1985) new series on the market value of federal debt outstanding into our government debt and real government interest variables. For these reasons, one should not proceed from Kormendi (1983), as did Feldstein and Elmendorf, but rather from the data definitions emerging from the 1986 debate.

Columns 3 through 5 of Table 2 show the effects of incorporating these data improvements. In column 3, we grant Feldstein and Elmendorf their choice of using only nominal data deflated by a uniform deflator. However, since consumption expenditures is not the dependent variable in either Feldstein and Elmendorf's regressions or ours, we use the deflator implied by the actual dependent variable, consumption flow. We then return to the real data on national product and its components, while deflating the financial variables by either the implicit consumption flow deflator (column 4) or the NNP deflator (column 5) as in Kormendi (1983). Except for granting Feldstein and Elmendorf their data on private wealth (a variable not at issue), column 5 returns to the data choices underlying Table 1.

In almost all cases, the results in columns 2 through 5 support the Consolidated Approach. In almost every case, the tax coefficients are insignificant, and in the vast majority of cases they are close to zero. Furthermore, there are as many positive tax coefficients as negative ones. In almost all cases, the government spending coefficients are negative and highly significant. Only the estimates in column 3 without the war reveal an insignificant effect of government spending. Even in these cases, however, Chow tests do not reject the hypothesis that the coefficients estimated without the war are equal to those for the full sample.[7]

The implications of Table 2 are easily summarized. Feldstein and Elmendorf's results obtain only as the *joint* consequence of (1) ignoring available real data, and (2) failing to incorporate state-of-the-art data and variable definitions. Had they chosen otherwise in either respect, they would have obtained results consistent with Kormendi (1983) and Kormendi and Meguire (1986). Moreover, and contrary to their claim, neither their results nor our own depend in any fundamental way on the inclusion or exclusion of data for World War II.

[6] Real data are not available for financial variables such as taxes and transfers. Hence these must be deflated by some deflator. Kormendi (1983) chose the NNP deflator out of consistency with his measure of income. The implicit deflator for consumption is a reasonable alternative.

[7] For example, the F-statistic for the homogeneity of the coefficients of the Basic regression is 1.69 with (6,39) degrees of freedom, which has a marginal significance level of 0.15.

TABLE 3—CONFIRMING THE TESTS OF KM (1986) WITH TEMPORARY TAXES

	F-Statistic	*F*-Statistic with
	from	Temporary Taxes (*TT*)
	KM (1986)	Included in Regression
	(*DOF*)	(*DOF*)
Test	*MSL*	*MSL*
1. Are Net Tax and Consistency	5.9	6.8
Restrictions Accepted Over	(6,20)	(6,18)
1952–84? (KM, Table 2)	0.001	< 0.001
2. Adding 1933–51 to:	1952–83	1952–85
a. PDL/Levels With Net Tax and	6.8	12.6
Consistency Restrictions	(19,24)	(19,24)
(KM, Table 3)	< 0.001	< 0.0001
b. PDL/Levels Without Net Tax and	4.3	3.3
Consistency Restrictions	(19,18)	(19,18)
(KM, Table 4, col. 1)	0.002	0.007
3. Adding 1931–51 to the Augmented	1.1	1.55
Consolidated Specification Estimated	(21,22)	(21,23)
in Differences (KM, Table 4, col. 2)	0.41	0.15

Note: MS = Modigliani and Sterling (1990). KM = Kormendi and Meguire (1986). Under the null, the statistics shown have F sampling distributions with (numerator, denominator) degrees of freedom (*DOF*). *MSL* = marginal significance level. *PDL* = first-degree polynomial distributed lags. The regressions underlying (2a) and (2b) differ from (1) and (4) in Table 1 of MS only in that they were estimated using the data described under "Data for Tables 3 and 4" in the Data Appendix. The regressions underlying (3) are similar to (1.1) in our Table 1.

II. Modigliani and Sterling

In our 1986 paper, we showed how to nest the restrictions implicit in Modigliani and Sterling's (MS) concept of "net taxes" (i.e., total taxes net of transfer payments and real government interest) into an unrestricted specification and so test the validity of the MS restrictions. We then established three findings: (1) granting Modigliani and Sterling their sample period and specification (polynomial distributed lag over levels data, henceforth "PDL/levels"), their implicit restrictions were rejected by their own data; (2) granting Modigliani and Sterling their PDL/levels specification but extending the sample period back to 1933 (using variable definitions as close to theirs as possible), their specification failed a test of homogeneity over the full period, whether or not the MS restrictions were imposed; and (3) in contrast, Kormendi's Augmented Consolidated specification was temporally homogeneous, yielding coefficients for government spending and taxation that rejected the MS specification in favor of the consolidated Approach.

In their rejoinder, Modigliani and Sterling (1990) claim that "[f]ailure to take temporary taxes into account biases the results against the Life Cycle Hypothesis." In Table 3, we present the F-statistics for the four tests undertaken in Kormendi and Meguire (1986). Column 1 contains for reference the results obtained in our earlier paper. Column 2 contains the same tests with "temporary taxes" added to the regressions. The inclusion of "temporary taxes" has no material effect on these tests. The net tax and disposable income restrictions in MS's preferred specification are still rejected. Whether or not these restrictions are imposed, the MS specification still fails a test of homogeneity when the data are extended back to 1933. Meanwhile, the data still support the homogeneity of the Augmented Consolidated specification, although not as strongly as before.

In Table 4, we present the government spending and tax coefficients from the eight

TABLE 4—GOVERNMENT SPENDING (*GS*) AND TAX (*TX*) COEFFICIENTS WITH AND WITHOUT TEMPORARY TAXES (*TT*): *MS* NET TAX AND CONSISTENCY RESTRICTIONS RELEASED

Number	Sample	*GS*	*TX*	*TT*
MS Specification: Levels Data, Polynomial Lags				
(1)	1952–84	−0.32 (0.20)	−0.12 (0.19)	–
(1′)	1952–84	−0.03 (0.23)	−0.63 (0.32)	0.60 (0.37)
(2)	1933–85	−0.17 (0.06)	0.23 (0.26)	–
(2′)	1933–85	−0.24 (0.04)	0.29 (0.20)	1.69 (0.44)
KM Specification: Differenced Data, No Lags				
(3)	1952–85	−0.26 (0.14)	0.00 (0.16)	–
(3′)	1952–85	−0.27 (0.13)	−0.16 (0.16)	0.69 (0.31)
(4)	1931–85	−0.26 (0.03)	0.06 (0.10)	–
(4′)	1931–85	−0.25 (0.03)	0.01 (0.10)	0.62 (0.28)

Note: KM = Kormendi and Meguire (1986). MS = Modigliani and Sterling (1986). *TT* for 1931–47 and 1985 is 0. When *TT* appears in a regression, *TX* is *net* of *TT*. Data are as follows: (1) and (1′), MS Table 3; (2) and (2′), same as those used to estimate col. 3 of Table 2 except for the addition of *EQ*; (3) through (4′), same as those used to estimate Table 1. Also see "Data for Tables 3 and 4" in the Data Appendix.

specifications that embody the three essential differences between us and Modigliani and Sterling. These coefficients are from unrestricted specifications estimated (1) with and without the period 1931–51, (2) in PDL/levels versus in differences without lags, and (3) with and without "temporary taxes." With one exception, all specifications yield significant negative government spending coefficients and insignificant tax coefficients (of mixed sign), results fully consistent with the Consolidated Approach. The one exception, (1′), is essentially regression (4) in Modigliani and Sterling.

But Modigliani and Sterling reject (4) as unreasonable a priori, because the effect of "temporary taxes" on consumption is positive and "improbably high." We fully concur, noting that the coefficients on temporary taxes (*TT*) in Table 4 are *always* large and positive. In the face of these positive coefficients, however, Modigliani and Sterling argue for a return to *restricted* specifications such as (1) and (2) in their Table 1, in which the net tax restriction works to "tame" the temporary tax coefficient so that its estimate is (in their judgment) reasonable. But, as we show in Table 3, whether "temporary taxes" are included or not, the MS restrictions are resoundingly rejected by the data.

Thus, we prefer to remain agnostic as to what, if anything, their "temporary tax" variable is capturing and reemphasize the results in Tables 3 and 4.[8] Seven of the eight

[8] While we area in principle sympathetic to a temporary/permanent decomposition of taxes and other variables (for example, Seater and Mariano, 1985), Modigliani and Sterling's temporary tax variable (*TT*) clearly does not yield reasonable results. This may stem from certain problematic aspects of the way they measure *TT*. First, note that temporary taxes $(TT)=0$ except for the years 1968–70 and 1975. In this regard, *TT* does not include the tax on excess corporate profits that was in force until 1954. Second, during much of the period 1968–70, when *TT* was positive, the business cycle was at a peak. Similarly, 1975 was a recession year and *TT* was negative. Hence *TT* possibly proxies for procyclical determinants of consumption not captured by the other explanatory variables. Third, *TT* is an estimate (which we were unable to replicate from original sources), derived under static assumptions, of the temporary component of federal income taxes alone for

unrestricted specifications shown in Table 4 firmly reject Modigliani and Sterling's prediction of significant negative tax effects and zero government spending effects. The one unrestricted case that supports their prediction is not, as they claim, due solely to including temporary taxes in the regression. Rather, it is the joint consequence of temporary taxes, a postwar sample, and PDLs applied to levels data. When any of these three specification choices are relaxed, their results fail to hold. Moreover, even in the one case that appears to support their prediction, the temporary tax coefficient is "implausible," causing Modigliani and Sterling to retreat to their restricted specification, which we have shown to be firmly rejected by the data. We do not believe this evidence supports their position.

III. Cointegration

Both of our critics take exception to our preference for a differenced specification. The substantive question underlying the choice of levels or differences for estimating a time series regression is whether its variables are cointegrated. We propose to resolve this matter by the following Monte Carlo exercise, which extends the methods of Engle and Granger (1987) and may prove of independent interest.

We simulate (1.1) under both cointegration and noncointegration so that our inferences do not depend on which of the two hypotheses is taken as the null.[9] We take the regressors as given,[10] and therefore only simulate the dependent variable *PC* under each hypothesis. To simulate *PC* under cointegration, we add AR(1) disturbances generated from iid Gaussian noise to the fitted values from (1.1) estimated in levels form with a maximum likelihood AR(1) correction. We simulate *PC* under noncointegration in a similar way, except that we use the fitted values from (1.1) estimated in differenced form, again with an AR(1) correction, and cumulate the resulting simulated changes in *PC* to obtain a simulated levels series. We then regress (OLS) each simulated *PC* series on the actual regressors, both in level and differenced form and record the resulting Durbin-Watson statistic. Finally, we draw inferences by comparing the Durbin-Watson statistic computed from the actual data to the simulated sampling distributions.

The results of this simulation exercise are shown in Table 5, which shows that the observed Durbin-Watson value of 0.69 computed from the level data is close to the simulated median under noncointegration (0.63). In contrast, 0.69 is fairly far in the left tail of the sampling distribution under cointegration, rejecting the null at the 10 percent level. Hence the observed Durbin-Watson value is much more consistent with noncointegration. Even more striking, the observed Durbin-Watson value of 1.34 derived from the differenced specification is fully consistent with noncointegration ($p = 0.62$), and inconsistent with cointegration ($p = 0.001$).

To summarize, we find no evidence that the variables in (1.1) are cointegrated. This result contrasts with empirical findings (surveyed in Stock and Watson, 1988) that income and consumption are cointegrated. Such findings are typically derived from specifications that use (a) disposable income

the years 1968–1970 and 1975. Even granting that the specific legislation in question was *intended* to be temporary, it is far from clear that there were no countervailing and offsetting factors in the overall budget process. Did agents really expect that revenues from the income tax surcharge of 1968 would not be followed by some other "revenue enhancing" legislation when it expired? Did the 1975 income tax "rebate" in effect offset the windfall profits tax on oil, also enacted around 1975? Whatever the answer to these questions, we cannot seriously entertain a positive coefficient for "temporary taxes" as measuring the effect on consumption of any tax.

[9] For a discussion of the importance of simulating the distribution of test statistics under both the null and the alternative in the context of testing for a unit root in real GNP, see Kormendi and Meguire (1990).

[10] We omit retained earnings (RE) from the simulated regressions because RE appears in the Augmented Consolidated specification only through the definition of disposable personal income. The RE coefficient in (1.1) is also insignificant and uncontroversial.

TABLE 5—SIMULATED DISTRIBUTION OF DURBIN-WATSON STATISTIC (DW) UNDER COINTEGRATION AND NONCOINTEGRATION.

Augmented Consolidated Specification
Sample Period 1930–85
2000 Monte Carlo Replications

Estimation	Actual		Percentiles of Simulated Distribution								
Method	DW	P	Min	5	10	25	50	75	90	95	Max
			Simulate Under Cointegration								
Levels	0.69	0.10	0.25	0.62	0.69	0.80	0.93	1.08	1.24	1.34	1.88
Differences	1.34	0.001	1.26	1.64	1.74	1.86	2.03	2.19	2.34	2.43	2.82
			Simulate Under Noncointegration								
Levels	0.69	0.59	0.24	0.35	0.40	0.50	0.63	0.80	0.95	1.05	1.54
Differences	1.34	0.62	0.59	0.88	0.96	1.10	1.27	1.45	1.63	1.71	2.23

Note: P = fraction of simulated values < actual DW. Low values of DW favor noncointegration. "Actual DW" is estimated from OLS run on the actual data. The simulated DW are from OLS estimates of (1.1) in levels as well as in differences, but with simulated instead of actual consumption and without retained earnings. To simulate consumption, we first estimate (1.1) over the actual data (in levels and in differences) with a maximum likelihood correction for AR(1) residuals, and recover the fitted values, ρ and σ (standard error of the estimate) from the estimate. We then add $z_t = \rho z_{t-1}\varepsilon_t$ to the fitted values of these regressions, where $\varepsilon_t \approx N(0,\sigma)$, with $(\rho,\sigma) = (0.798, 0.0299)$ under cointegration, and (0.492, 0.0271) under noncointegration. The first simulated value of level consumption is set equal to its actual value in 1930. Each simulated z_t series is 155 observations long, with the first 100 values being discarded to eliminate the effect of an arbitrary starting value of 0. All calculations were performed using version 3.0 of *RATS*.

instead of unrestricted fiscal variables, and (b) postwar data only, rather than the longer data starting in 1929. Contrary to the claims of our critics, our use of a differenced specification is supported by the data.[11]

IV. Conclusion

Feldstein and Elmendorf's results obtain only as the joint result of (a) not using available real data on consumption, income, and government spending, and (b) failing to implement the improved variable definitions that emerged from the 1986 debate. Moreover, and contrary to their claim, neither their results nor ours depend in any essential way on whether data for the period 1941–46 are included in the sample. None of our 1986 findings are overturned in any way by including Modigliani and Sterling's proposed measure of temporary taxes in the regressions. Furthermore, the coefficient of temporary taxes in unrestricted specifications is large and positive. Finally, simulation tests extending Engle and Granger (1987) show that our Augmented Consolidated specification is not likely to be cointegrated, which supports our choice of estima-

[11] Modigliani and Sterling argue that differencing is inappropriate because their PDL/levels estimates "...reject the hypothesis that rho could...equal 1, the value implicit in the differenced specification." However, the sampling distribution of test statistics for cointegration, including ρ and the Durbin-Watson statistic (DW), are not known when these are computed from multiple regressions with PDL variables, estimated over small samples. In particular, contrary to Modigliani and Sterling's claim, neither the simulations nor the results of Engle and Granger (1987) apply here because these were derived for bivariate regressions of nondurable consumption on disposable income alone. Moreover, the negative first-order autocorrelation in Modigliani and Sterling's equation (4) is an artifact of their use of PDLs and the small number of degree of freedom. A Monte Carlo exercise (available upon request) analogous to Table 5 shows that adding simulated white noise to the fitted values of (4) and estimating with PDL/levels yields a mean DW of 2.86, the value obtained from an OLS estimate of (4). This perhaps surprising result is consistent with Johnston (1984, equation (8–63)), who shows that given n observations, K regressors (with each PDL variable yielding 2 regressors) and white noise disturbances, $E(DW) \cong 2[(n-1)/(n-k)]$, or 3.56 in the case of (4). This negative autocorrelation is also consistent with the overfitting suggested by the failure of (4) to validate out-of-sample (Table 3, row 2b).

tion in differences. The results in Kormendi (1983) and Kormendi and Meguire (1986), that is, significant negative effects of government spending with no significant effects of taxes, remain robust to the challenges posed by the critics.

After two rounds of debate, we wish to highlight certain methodological contributions of the Consolidated Approach. Under what we term the Standard Approach, consumption is primarily a function of disposable personal income, that is, $Y - TX + TR + RGINT - RE$. This approach restricts the effect on consumption of the components of disposable income to be the same (except for sign) and allows no role for government spending. Under the pure Ricardian Approach, consumption is a function of "total disposable income," that is, $Y - GS$, where government spending is assumed to represent the true resource burden of the government sector on the private sector. Under the Consolidated Approach, consumption is a function of Y, GS, and TR, each with possibly distinct effects, thus allowing government spending to substitute for private consumption and government redistribution of income to affect the aggregate propensity to consume.

The key to the tests developed in Kormendi (1983) was to nest these and other hypotheses by releasing all restrictions on the fiscal variables in the consumption function. Doing so yields the Augmented Consolidated Approach, in which consumption is a function of Y, GS, TX, TR, and $RGINT$. This approach is equivalent to including the right-hand-side variables from the government's budget constraint (Deficit $= TX - GS - TR - RGINT$) as explanatory variables in the consumption function. The estimated coefficients of the fiscal variables can then be interpreted as the effects on consumption of deficit-financed marginal changes in these variables.[12]

The importance of releasing implicit restrictions on the fiscal variables transcends the particular time series aggregate consumption function context in which it was originally proposed. In virtually any context in which the relation between private consumption and income is addressed, whether it be Euler equations, cointegration tests, vector autoregressions, or tests of the sensitivity of consumption to income innovations, one can no longer use disposable personal income without question. At minimum, one must test, using an unrestricted specification, whether aggregate income, government spending, taxes, and transfers each have distinct effects on consumption.

DATA APPENDIX

This appendix updates and extends the Data Appendix in Barth, Iden, and Russek (1986) (henceforth BIR) and should be read in conjunction with it. In particular, it defines (a) the variables used in (1.4) and the rows labeled "federal only" in Table 2, and (b) new measures of the market value of government debt and of real interest paid by the government. All data are rounded to the nearest \$U.S. 100 million, and real data are rescaled into 1972 prices. As POP is rounded to the nearest 100,000, all variables are denominated in thousands of \$U.S. per capita at 1972 prices. Unless otherwise stated, all data are in current prices, and all outstandings are measured as of the end of the preceding year. Data series or variable names ending in "F" (real) or "FN" (nominal) are measured only over the federal government.

The following abbreviated references are used below:

BS	*Business Statistics*, a biannual supplement to the *SCB*.
ERP	*1988 Economic Report of the President to the Congress*.
FF	*Flow of Funds Accounts, Financial Assets, and Liabilities, Year-end 1963–86* (Z.1), Table "State and Local Government—General Funds."
FRTW	*Fixed Reproducible Tangible Wealth in the US: 1925–85.*
NIPA	*National Income and Product Accounts of the United States: Historical Tables, 1929–82*, updated to 1985 using the July 1987 *SCB*. The notation "$x.y.z$" refers to line number z in *NIPA* Table $x.y$.
SA	*Statistical Abstract of the United States*, 1984 issue.
SCB	*Survey of Current Business*.

Data Differing from BIR only by Updates and Revisions: *CDR*, *CNDR*, *CSR*, *DURR*, *GSR*, *NNPR*, *PNNP* (latter three variables now *NIPA* series 1.2.18, 1.9.5, and 7.7.3), *REN*, *TRN*, and *TXN* are as defined in BIR, except revised and updated as per *NIPA*. Population

[12] See Charette (1986) for further tests of the stock-flow aspects of the Government Budget Constraint in the context of Hayashi's (1982) model applied to Canadian data.

(POP) for 1928–38 is from SA, Table 2, column "Resident Population," and for 1939–85, from ERP, Table B-30, col. 1. Data for KN, the total net stock of fixed private capital used in the computation of W, are from $FRTW$, Table A13, p. 243, col. 2, with updates for 1984–85 from SCB, August 1987, p. 100, Table 2, row 1.

Data for Deflators other than PNNP: Let CDN, $CNDN$, CSN ($NIPA$, 1.1.3–5) be consumption expenditures in current prices on durables, nondurables, and services. Let $DURN$ ($FRTW$, Table A18, Net stock, "Total, all types," col. 2, with updates from August 1987 SCB, Table 20, col. 1) be the net stock of consumer durables, likewise in current prices. Then

$$PPC = (0.3CNDN + 0.1CDN + CNDN + CSN)/ (0.3CNDR + 0.1CDR + CNDR + CSR).$$

Also,

$PPCX$ = Implicit Deflator for Personal Consumption Expenditures ($NIPA$, 7.4.2)

Feldstein and Elmendorf's Data: Actual values for GB, GBF, PC, W, and $FEPOP$ (resident population) kindly provided by Douglas Elmendorf. We rescale $FEPOP$ into 100,000s, and the other data into $1000 at 1972 prices. Unlike POP, $FEPOP$ does not include Armed Forces personnel stationed overseas.

Data for Tables 3 and 4

1952–84: Table 3, Modigliani and Sterling (1986). Data for government interest ($RGINT$) and the 1948 values of all variables were kindly provided by Arlie Sterling.

1931–85: The correspondence between variables defined in this Appendix (KM) and the variables appearing in Modigliani and Sterling (1990) (MS) is as follows:

MS:	A	C	D	E	G	$RGINT$	T	T^*	TR
KM:	W, EQ	PC	DEF	GS	GB	$RGINT$	$TX - T^*$	TT	TR

where

DEF $= GS + TR + RGINT - TX$

EQ $= MVLSN/SDIVI$

$MVLSN$ = Market value at year-end of stocks listed on the New York Stock Exchange (BS, various editions, esp. 1967 (p. 108) and 1986 (p. 77), far right column on page).

$SDIVI$ is defined in the Note to Table A1.

TT $= 0$ (1930–67, 1971–74, 1976–85);
$= T^*$ from Table 3 in MS (1968–70, 1975).

When TT is included in the regression, TX is net of TT.

Data for GB, Market Value of Government Debt

Data with sources other than FF are December values for the preceding year. $GBN3$ (through 1976) and $GBN1$ are exactly as described in BIR.

$CMISLN$ = Par Value of Credit Market Liabilities Issued by State and Local Governments (FF, row 13).

$GBN2$ = Market Value of Privately Held Gross Federal Debt (Cox, 1985, Table 2).

MBP = Average Market Price of Municipal Bonds per Dollar of Principal, 1977–85 (data courtesy of Douglas Elmendorf).

$SHLTEON$ = Par Value of State and Local Holdings of Tax-Exempt Obligations (FF, row 9). Assuming these securities to be mostly short term, then market and par values are approximately equal.

$SLBPN$ = Par Value of Long Term Tax-Exempt State and Local Debt (FF, row 13).

$GBN4$ $= CMISLN - SLBPN(1 - MBP) - SHLTEON$ (1976–85)

$GBN3$ $= GBN4(GBN3_{76}/GBN4_{76})$ (1977–85)

GBN $= GBN2 + GBN3$ (1943–85)
$= GBN1(GBN_{43}/GBN1_{43})$ (1929–42)

Data for Government Interest, RGINT

$FRPROFN$ = Federal Reserve Bank Profits paid to the U.S. Treasury ($NIPA$, 3.2.7).

$GDIV$ = Corporate Dividends received by Government ($NIPA$, 3.1.18).

$GINTN$ = Net Interest paid by Government ($NIPA$, 3.1.13).

$GINTFN$ = Net Interest paid by Federal Government ($NIPA$, 3.2.22).

$GINTFORN$ = Government Interest paid to Foreigners ($NIPA$, 3.1.16).

P_t = Value in period t of whichever of $PPCX$, PPC or $PNNP$ we use to construct $DIVI$.

PI_t $= \ln((P_{t+1} + P_t)/(P_t + P_{t-1}))$

DRB $= PI \times GBN$

$DRBF$ $= PI(GBN - GBN3)$

Other Fiscal Data

GSN = Government Purchases of Goods and Services ($NIPA$, 1.1.18).

$GSSLN$ = State and Local Government Purchases ($NIPA$, 1.1.22).

$GSSLR$ = State and Local Government Purchases in constant prices ($NIPA$, 1.2.22).

$NSUBN$ = Government Subsidies to Businesses Net of Current Surplus of Government Enterprises ($NIPA$, 3.1.19).

$NSUBFN$ = That part of $NSUBN$ attributable to the federal government ($NIPA$, 3.2.27).

$TRFN$ = Federal Government Transfer Payments to Domestic Persons ($NIPA$, 3.2.19).

$TRFORN$ = Government Transfer Payments to Foreigners ($NIPA$, 3.1.12).

$TXFN$ = Federal Tax Collections ($NIPA$, 3.2.1).

TABLE A1—SUMMARY DEFINITIONS OF VARIABLES USED IN TABLES 1 AND 2

	Column Number from Table 2			
Variable	(1)	(2)	(3)	(4)
$DIVI$	$PPCX \times FEPOP$	Same	$PPC \times POP$	Same
GB	FE	Same	$GBN/SDIVI$	Same
GBF	FE	Same	$\frac{GBN - GBN3}{SDIVI}$	Same
$RGINTF$	$\frac{GINTN\#}{DIVI}$	Same	$\frac{GINTN\# - FRPROFN - DRB - GDIV}{DIVI}$	Same
$RGINTF$	$\frac{GINTFN\#}{DIVI}$	Same	$\frac{GINTFN\# - FRPROFN - DRBF}{DIVI}$	Same
GS	$GSN/DIVI$	$GSR/FEPOP$	$(GSN/DIVI) + GS\#$	$(GSR/POP) + GS\#$
$GSSL^*$	$GSSLN/DIVI$	$GSSLR/FEPOP$	$GSSLN/DIVI$	$GSSLR/POP$
PC	$(0.3DURN + 0.1CDN + CNDN + CSN)/DIVI$	$PCR/FEPOP$	PCR/POP	Same
TR	$TRN/DIVI$	Same	$(TRN + NSUBN)/DIVI$	Same
TRF	$TRFN/DIVI$	Same	$(TRFN + NSUBFN)/DIVI$	Same
TX	$TXN/DIVI$	Same	$(TXN - FRPROFN)/DIVI$	Same
TXF	$TXFN/DIVI$	Same	$(TXFN - FRPROFN)/DIVI$	Same
Y	$NNPN/DIVI$	$NNPR/FEPOP$	$NNPN/DIVI$	$NNPR/POP$

$^*GSF = GS - GSSL$

Note: The variable definitions under (4) also hold for col. 5 and all of Table 1, except that $DIVI = PNNP \times POP$, and W in Table 1 is defined by (A8) in BIR. RE is always $REN/DIVI$. Deflators are normalized so that 1972 = 1.

Same = same formula as the entry immediately to the left, except that $DIVI$ ($SDIVI$) is taken from the column in which "Same" appears.

FE = Feldstein and Elmendorf's actual data, rescaled to a 1972 base; also the case for PC in (1), for $FEPOP$ in (1) and (2), and for W in Table 2.

$GINTN\#$ $= GINTN - GINTFORN$

$GINTFN\#$ $= GINTFN - GINTFORN$

$GSN\#$ $= (TRFORN + GINTFORN)/DIVI$

$NNPN$ = Net National Product ($NIPA$, 1.9.5).

PCR $= 0.3DURR + 0.1CDR + CNDR + CSR$.

$SDIVI_t$ $= (DIVI_t - DIVI_{t-1})/2$.

REFERENCES

Ando, Albert and Modigliani, Franco, "The 'Life-Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *American Economic Review*, March 1963, *53*, 55–84.

Barth, James, Iden, George and Russek, Frank S., "Government Debt, Government Spending, and Private Sector Behavior: Comment," *American Economic Review*, December 1986, *76*, 1158–67.

Brunner, Karl, "Fiscal Policy in Macro Theory: A Survey and Evaluation," in *The Monetary Versus Fiscal Policy Debate*, R. W. Hafer, ed., London: Rowman and Allanheld, 1986.

Charette, Michael F., "Canadian Tests of Direct Crowding Out Effects on Aggregate Consumption," University of Windsor Working Paper, 1986.

Cox, W. Michael, "The Behavior of Treasury Securities: Monthly, 1942–1984," *Journal of Monetary Economics*, September 1985, *16*, 227–40.

Engle, Robert and Granger, C. W. J., "Co-integration and Error Correction: Representation, Estimation and Testing," *Econometrica*, March 1987, *55*, 251–76.

Feldstein, Martin, "Government Deficits and Aggregate Demand," *Journal of Monetary Economics*, January 1982, *9*, 1–20.

______ **and Elmendorf, Douglas,** "Government

Debt, Government Spending and Private Sector Behavior: Comment," *American Economic Review*, June 1990, *80*, 589–99.

Hayashi, Fumio, "The Permanent Income Hypothesis: Estimation and Testing by Instrumental Variables," *Journal of Political Economy*, October 1982, *90*, 895–916.

Johnston, John, *Econometric Methods* (3rd ed.). New York: McGraw-Hill, 1984.

Kormendi, Roger C., "Government Debt, Government Spending and Private Sector Behavior," *American Economic Review*, December 1983, *73*, 994–1010.

______ **and Meguire, Philip,** "Government Debt, Government Spending and Private Sector Behavior: Reply," *American Economic Review*, December 1986, *76*, 1180–87.

______ **and** ______, "The Government Budget Constraint, Income Redistribution and Private Saving," University of Michigan Working Paper, October 1989.

______ **and** ______, "A Multicountry Characterization of the Nonstationarity of Aggregate Output," *Journal of Money, Credit, and Banking*, February 1990, *22*.

Modigliani, Franco and Sterling, Arlie, "Government Debt, Government Spending, and Private Sector Behavior: Comment," *American Economic Review*, December 1986, *76*, 1168–79.

______ **and** ______, "Government Debt, Government Spending and Private Sector Behavior: A Further Comment," *American Economic Review*, June 1990, *80*, 600–603.

Seater, John J. and Mariano, Roberto S., "New Tests of the Life Cycle and Tax Discounting Hypotheses," *Journal of Monetary Economics*, March 1985, *15*, 195–215.

Stock, James and Watson, Mark, "Variable Trends in Economic Time Series," *Journal of Economic Perspectives*, Summer 1988, *2*, 147–74.

Board of Governors of the U.S. Federal Reserve System, Flow of Funds Section, *Flow of Funds Accounts, Financial Assets and Liabilities, Year-end 1963–86* (*Z.1*), Washington: 1988.

Economic Report of the President to the Congress, Washington: USGPO, 1988.

Statistical Abstract of the United States, Washington: USGPO, 1984.

U.S. Department of Commerce, *National Income and Product Accounts of the United States: Historical Tables, 1929–82*. Washington: USGPO, 1986.

______, *Business Statistics*. Washington: USGPO, 1986.

______, *Fixed Reproducible Tangible Wealth in the US: 1925–85*. Washington: USGPO, 1987.

Tobin's *q* and the Structure-Performance Relationship: Comment

By JERRY L. STEVENS*

The exchange of comments between William Shepherd and Michael Smirlock, Thomas Gilligan, and William Marshall (this *Review*, December 1986) raised two key points that remain unresolved. The first point is whether Tobin's q ratio, a firm's financial market value divided by replacement cost of its assets, is a better measure of firm performance than accounting rates of return. The second point of contention is whether superior performance, however measured, can be attributed to efficiency rather than market power. This paper offers further clarification on both of these points.

In Section I the performance measure choice is shown to be influenced by fundamental differences between finance and economics. In Section II, the structure-performance model employed by Smirlock, Gilligan, and Marshall (hereafter, SGM) and Shepherd is shown to be a special case of a more general model allowing for a dependence of the market-share-performance relationship on the concentration ratio. The same data from the original study by SGM (1984) are used in Section III to provide a comparison of SGM's findings with empirical results from an alternative specification of the structure-performance model. This comparison suggests that attributing superior firm performance exclusively to efficiency is not well founded. Concluding remarks are found in Section IV.

I. Tobin's *q* and Accounting Rates of Return

Shepherd and SGM debated the merits of Tobin's q ratio as an alternative to more traditional accounting rates of return, such as net income to sales, as a measure of firm performance.[1] Objections to accounting rates of return have focused on accounting measurement errors from arbitrary depreciation schedules and expensing of intangible assets (for example, advertising and research and development).[2] The truth is that both q and accounting rates of return are subject to many of the same measurement problems to some degree. To test the sensitivity of q and accounting rates of return to measurement errors, Henry McFarland (1988) used Monte Carlo experiments to determine which accounting measure provides the best approximation to its "true" measure and found that q estimates have smaller average errors than accounting rate of return measures. In addition, the q ratio was found to have a much higher average correlation with its true measure.

A fundamental issue in the q versus accounting return debate is that profitability and value are often used interchangeably as performance standards even though they each represent a different phenomenon. Much of the disagreement on this point reflects a difference between finance and economics perspectives. Research questions for economists tend to revolve around profitability and price-costs measures, while re-

*Associate Professor of Finance, E. C. Robins School of Business, University of Richmond, VA 23173. I am indebted to William Shepherd for providing data and helpful advice. I also thank Manuel Jose and Len Nichols for their comments and encouragement.

[1] Empirical studies in industrial organization tend to measure performance with net income divided by sales. Earnings before interest and taxes would be more appropriate than net income for price-cost margins studies since differences in financing, depreciation, and tax effects unrelated to pricing and efficiency of production affect net income.

[2] For a discussion of accounting rates of return as measures of firm performance, see Franklin M. Fisher and John J. McGowan (1983), Ira Horwitz (1984), William F. Long and David J. Ravenscraft (1984), Michael van Breda (1984), and Franklin M. Fisher (1984).

search questions in finance tend to revolve around valuation, where value is determined by what investors are willing to pay for claims against the firm.[3] Shepherd has argued for accounting rates of return, consistent with the traditional concern in industrial organization economics for the firm's ability to extract higher price-cost margins from sales. SGM, trained in finance, advocate Tobin's q ratio as a performance measure since q reflects *ex ante* financial market valuation of all rents to the firm including, but not limited to, both the level and risk of future profitability.

Shepherd found q ratios to be conceptually debatable because, unlike accounting profit, q is a phenomenon of capital market valuation, not of the firm. But in finance, the objective of the firm's management is to maximize financial market valuation of the firm. Extensive documentation of the semistrong form of efficient security pricing suggests that accounting data is fully reflected in stock price valuation.[4] Clearly, the debate is driven by differences in objectives represented by profits as opposed to value.

While the debate tends to break down along disciplines, there are reasons why industrial organization economists might use q in market power studies. Since market structure has been shown to affect the risk of cash flows (see, for example, Marti Subrahmanyam and Starvos Thomadakis (1980) and Manuel L. Jose and Jerry L. Stevens (1987)), the structure-q relationship represents a broader set of market power influences. For example, a firm may use market power to lower risk rather than increase the price-cost margin. If so, the market value of the firm would be enhanced even though there was no significant relationship between market power and accounting rates of return for that period.[5] Measures of q also discriminate well on the basis of profitability. Using Monte Carlo techniques McFarland (1988) found that q estimates were neither consistently better nor consistently worse than accounting rates of return in detecting supracompetitive profits.[6]

II. Specification of the Structure-Performance Hypotheses

The following regression model was used by SGM and a number of prior researchers, including Shepherd (1972) and Bradley T. Gale and Ben S. Branch (1982), to analyze performance of the ith firm:

$$(1)\quad \text{Performance}_i = a + b\text{MS}_i + c\text{CR}_i + d\text{MSG}_i + e\text{MBE}_i + f\text{HBE}_i + d\text{error term}_i,$$

where MS represents firm market share, CR is the four-firm concentration ratio, MSG is market share growth, MBE is a medium-barrier to entry dummy variable, and HBE is a high-barrier to entry dummy variable. SGM found that market share dominates the concentration ratio (b is significant and high compared to c) and used this finding to conclude that "efficiency," rather than "collusion," is responsible for superior performance.

Shepherd (1986) questioned SGM's conclusion that a finding of significant positive market-share-q and insignificant concentration-q relationships reflected cost efficiency rather than price collusion.[7] Shepherd ar-

[3] These differences between finance and economics are discussed by Lawrence Summers (1985).

[4] For a reference to empirical verification of semistrong efficient market pricing, see Thomas E. Copeland and J. Fred Weston (1983).

[5] Richard Schramm and Roger Sherman (1974) offer a more complete discussion of how market power can increase the value of the firm through lower risk rather than higher profits. Also, see William G. Schwert (1981) for a discussion of the potential use of financial valuation concepts in economics.

[6] Shepherd and SGM both found that substituting q for accounting profits did not change the basic findings of market share dominance.

[7] The SGM assumption that market share's influence on firm performance is due exclusively to efficiency is only one of three different interpretations found in the literature. At the other extreme, Stephen Rhoades (1985) explains positive market-share-performance findings as

gued that concentration is not the exclusive form of market power and that market share "dominance" advantages affect prices. Unfortunately, Shepherd's attention was diverted away from modeling the joint influence of market share and concentration to a concern for collinearity between market share and concentration. SGM's (1986) response showed that their results were not sensitive to collinearity.

Rather than focus on collinearity between market share and market concentration, the sensitivity of SGM's findings to the specification of the market-share-q relationship should be tested. SGM's model specified a constant market share coefficient as if a firm's market share has the same relationship with q in a highly concentrated market as in a less concentrated market. An alternative specification of equation (1) that allows for an interaction of market share and concentration provides a test for the assumption that the concentration-q relationship is captured exclusively by the concentration ratio coefficient of equation (1).

There is good reason to expect a dependence of market-share-q relationships on market concentration. Dennis C. Mueller (1986) has argued that market dominance advantages for a firm depend on the firm's market share relative to the distribution of shares by its rivals. For example, if market concentration offers price collusion, inefficient firms survive under the pricing umbrella and higher profitability will correspond with higher market shares. In this case both collusion and efficiency compete to explain the same positive market-share-q phenomenon even if they are not collinear. Dependence of the market share-performance relationship on concentration $\partial q/\partial \mathrm{MS}$ depends on the level of CR. According to this "collusion" argument, the interaction coefficient would be positive reflecting higher premiums for market share when collusive pricing in a concentrated industry is present. However, more than one interaction hypothesis is plausible. An "efficiency" interaction hypothesis suggests that a firm with high market share attracts a premium due to efficiency, but that the premium will be lower if there is oligopolistic competition with other large market share firms. This view is consistent with a positive-market-share-q coefficient and a negative coefficient for the (MS × CR) interaction variable.

The "collusion" interaction hypothesis is consistent with Shepherd's objection to SGM's conclusions. Individual coefficients of market share and concentration in the SGM model would be biased due to the restriction of a zero interaction effect as both market share and concentration compete to explain the same phenomenon. The "efficiency" interaction hypothesis is that the market-share-q correspondence would show even stronger efficiency effects if market concentration's downward bias on the estimated relationship were recognized. These competing hypotheses are tested in the following model:

$$(2)\quad q_i = a + b\mathrm{MS}_i + c\mathrm{CR}_i + d\mathrm{MSG}_i + e\mathrm{MBE}_i + f\mathrm{HBE}_i + g(\mathrm{MS}\times\mathrm{CR})_i + \text{error term}_i,$$

where all variables are defined as before. The SGM specification is a special case of equation (2) where the correspondence between market share and firm performance is constrained to be a constant with respect to the concentration ratio. If the SGM specification is correct, the t-test for the interaction coefficient (g) will not be statistically significant. The collusion (efficiency) interaction hypothesis is supported by a statistically significant positive (negative) sign for the market-share-concentration (MS × CR) interaction coefficient.

rents from lower elasticity of demand due to inherent product differentiation. A middle ground is provided in the work of Roger Clarke, Stephen Davies, and Michael Waterson (1984), where a variety of alternative phenomena are used to explain market share-performance findings.

III. Reproduction of SGM's Empirical Work

To determine whether interaction effects alter SGM's conclusions, regression results for equation (1) were obtained from the same Shepherd (1972) data panel employed by SGM. Tobin's q was constructed for the firms in Shepherd's data using the procedures of Eric Lindenberg and Stephen Ross (hereafter, L&R) (1981) and SGM (1984).[8] To validate the q-program, constructed q values were compared to published q values from the L&R paper for the same firms over the same time period. Tests for differences in means and variances were conducted and the null hypothesis of equal means and variances could not be rejected. A correlation of a 0.98 existed between computed q values and the L&R values. These findings verified the consistency of procedures used to construct the q values of this study.

Table 1 provides a comparison of published regression results from the SGM paper for equation (1) and the identical model estimated from the reconstructed data set (RDS).[9] Results from the reconstructed data set are similar to the SGM findings with market share dominance over concentration when the traditional model of equation (1) is employed.

TABLE 1—SGM AND RECONSTRUCTED DATA SET (RDS) REGRESSION RESULTS[a]

	Equation (1)		Equation (2)
Independent Variables	SGM Results	RDS Results	RDS Results
Intercept	−2.600	−2.010	2.480
	(3.51)[b]	(1.93)[b]	(1.63)
MS	0.055	0.070	−0.165
	(3.61)[b]	(5.31)[b]	(−2.64)[b]
CR	0.009	0.013	−0.049
	(1.05)	(1.15)	(−2.52)[b]
MSG	2.670	1.270	1.060
	(4.50)[b]	(1.42)	(1.27)
MB	0.261	−0.046	0.429
	(0.90)	(−0.13)	(0.429)
HB	0.380	0.336	0.910
	(0.95)	(0.64)	(1.78)
(MS×CR)			0.003
			(3.84)[b]
R^2	NR[c]	0.437	0.517
F	17.06	13.99	15.88

[a] The dependent variable is Tobin's q ratio. t-statistics appear in parentheses.

[b] Coefficient significant at the 5 percent level.

[c] SGM did not report (NR) the R^2 values for their regressions.

Regression results for equation (2) are also reported in Table 1. The high level of significance for the positive interaction term supports the "collusion" hypothesis and misspecification of the SGM model.[10] A level of market concentration greater than 55 percent, compared to a sample mean of 65.14 percent, is required before a positive market-share-q correspondence is obtained ($\partial q/\partial \mathrm{MS} > 0$ if $-0.165 + 0.003\mathrm{CR} > 0$). Market concentration has a positive influence on q for firms with market shares above 16.33 percent below the market share sample mean of 23.51 percent ($\partial q/\partial \mathrm{CR} > 0$ if $-0.049 + 0.003\mathrm{MS} > 0$).[11]

Since multiple regressors appeared in the significance tests for concentration and market share, F-tests of vectors of coefficients were also conducted. The F-test results support the significance of both market share and concentration with a computed F of

[8] Values for q had to be constructed because SGM's data were not made available when requested. However, SGM also used the Lindenberg and Ross procedures so a validation of the q program was possible. Additional summary measures of the reconstructed data set and comparisons with the SGM data are available from the author upon request.

[9] The diagnostics described by David Belsley, Edward Kuh, and Robert Welsch (1980) were employed to evaluate Shepherd's concern over potential degrading collinearity between variables in the SGM model. Results from the PROC REG procedure of SAS (Statistical Analysis System software package, version 5) verified the SGM conclusion that degrading collinearity was not present.

[10] The interaction term (MS×CR) is collinear with MS and CR, causing the t-tests to be lower on the individual MS and CR coefficients. Even so, both MS and CR coefficients remain significant.

[11] David J. Ravenscraft (1983) employed an interaction term in his empirical work with line of business data. Differences in data and variable measurement prevent a comparison of Ravenscraft's results with those from Shepherd's data panel.

8.21 for the hypothesis that $\partial q/\partial \mathrm{CR}=0$ and a computed F of 23.76 for the hypothesis that $\partial q/\partial \mathrm{MS}=0$, both compared to a critical F value of 4.88 at the 1 percent level of significance. These results along with the significance of the interaction variable in Table 1 modify the SGM conclusions. Concentration is no longer "dominated" by market share since a statistically positive market-share-q relationship depends on a high level of concentration. It would be wrong to conclude that market share dominates when rents due to market share cannot be separated from rents due to concentration. The SGM "rents to efficiency" conclusion appears to rest on a misspecification of the structure-performance model.

IV. Conclusion

The controversy surrounding the findings and methods of Smirlock, Gilligan, and Marshall has led to a productive exchange of ideas. Differences between single-period profits and a financial valuation view of performance have been identified and additional paths of influence of market power on firm performance may now be investigated. For example, it would be interesting to know how the interaction between market share and concentration affects q through systematic risk. If firms with higher market share in less concentrated industries are less risky, lower q values for such firms would be due to lower returns rather than higher risk. Such research may narrow the gap between the views of firm performance in economics and finance.

While Smirlock, Gilligan, and Marshall presented their results as proof of the dominance of efficiency over collusion, the sensitivity of their results to the specification of the regression model has been demonstrated. Specifically, positive returns to market share have been shown to depend on a high level of market concentration, suggesting that collusion and efficiency explanations are not clearly separated. Attributing superior performance to efficiency alone is not substantiated given the interdependence found between the market-share-q relationship and the concentration ratio.

REFERENCES

Belsley, David, Kuh, Edward and Welsch, Robert, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: Wiley & Sons, 1980.

Clarke, Roger, Davies, Stephen and Waterson, Michael, "The Profitability-Concentration Relation: Market Power or Efficiency?" *Journal of Industrial Economics*, June 1984, *32*, 435–50.

Copeland, Thomas E. and Weston, J. Fred, *Financial Theory and Corporate Policy*, 2nd ed., Reading, MA: Addison Wesley, 1983, 317–53.

Fisher, Franklin M., "The Misuse of Accounting Rates of Return: Reply," *American Economic Review*, June 1984, *74*, 509–20.

______ and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, *73*, 82–97.

Gale, Bradley T. and Branch, Ben S., "Concentration Versus Market Share: Which Determines Performance and Why Does It Matter?" *Antitrust Bulletin*, Spring 1982, *27*, 83–105.

Horwitz, Ira, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 492–93.

Jose, Manuel L. and Stevens, Jerry L., "Product Market Structure, Capital Intensity, and Systematic Risk: Empirical Results from the Theory of the Firm," *Journal of Financial Research*, Summer 1987, *10*, 161–75.

Lindenberg, Eric and Ross, Stephen, "Tobin's q Ratio and Industrial Organization," *Journal of Business*, January 1981, *54*, 1–32.

Long, William F. and Ravenscraft, David J., "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 494–500.

McFarland, Henry, "Evaluating q as an Alternative to the Rate of Return in Measuring Profitability," *Review of Economics and Statistics*, November 1988, *70*, 614–22.

Mueller, Dennis C., *Profits in the Long Run*, Cambridge: Cambridge University Press, 1986.

Ravenscraft, David J., "Structure-Profit Relationships at the Line of Business and Industry Level," *Review of Economics and*

Statistics, February 1983, *65*, 22–32.

Rhoades, Stephen, "Market Share as a Source of Market Power: Implications and Some Evidence," *Journal of Economics and Business*, December 1985, *37*, 343–63.

Schramm, Richard and Sherman, Roger, "Profit Risk Management and the Theory of the Firm," *Southern Economic Journal*, January 1974, *40*, 353–63.

Schwert, William G., "Using Financial Data to Measure Effects of Regulation," *Journal of Law and Economics*, March 1981, *24*, 121–58.

Shepherd, William, G., "Tobin's *q* and the Structure-Performance Relationship: Comment," *American Economic Review*, December 1986, *76*, 1203–09.

______, "The Elements of Market Structure," *Review of Economics and Statistics*, February 1972, *54*, 25–37.

Smirlock, Michael, Gilligan, Thomas and Marshall, William, "Tobin's *q* and the Structure-Performance Relationship," *American Economic Review*, December 1984, *74*, 1051–60.

______, ______ **and** ______, "Tobin's *q* and the Structure-Performance Relationship: Reply," *American Economic Review*, December 1986, *76*, 1211–13.

Subrahmanyam, Marti and Thomadakis, Stavros, "Systematic Risk: The Theory of the Firm," *Quarterly Journal of Economics*, May 1980, *94*, 437–51.

Summers, Lawrence, "On Economics and Finance," *Journal of Finance*, July 1985, *40*, 633–35.

van Breda, Michael F., "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 507–17.

Cooperation, Harassment, and Involuntary Unemployment: Comment

By ERNST FEHR*

Recently A. Lindbeck and D. Snower published an interesting paper in the *American Economic Review* where they try to show that cooperation and harassment activities of insiders toward entrants cause an underbidding failure. This means that although the utility of a job is higher than the utility of unemployment, firms and unemployed workers are not able to sign a contract which stipulates that a particular job is performed at less than the prevailing wage. Since insiders have the power to withdraw cooperation from entrants, they are able to make them less productive. By harassment activities they increase the disutility of work and the reservation wage of entrants and make them more expensive. Withdrawal of cooperation and harassment reduces the (potential) profitability of unemployed workers and may allow insiders to enforce nonmarket clearing wages.

In this comment[1] I will show (i) that they apply incompatible assumptions. As a consequence *all* insiders are replaced by outsiders if they threaten to harass entrants and if they set their wages according to the wage rule of Lindbeck and Snower. (ii) If harassment activities h_E are utility decreasing for insiders, $(\delta\Omega/\delta h_E < 0)$ $h_E > 0$ is no credible threat and in equilibrium $h_E = 0$ prevails. (iii) There exist Pareto-improving contracts which eliminate harassment and induce insiders to cooperate with entrants. Hence, in equilibrium only voluntary unemployment prevails.

If the number of insiders (m in each firm) is so large that their marginal product (Af') is below their reservation wage $R_I = 1$ (scenario 1), the work force is reduced until $Af' = R_I$ (see Figure 1). In this scenario unemployed workers are never profitable for the firm even if full cooperation and no harassment were to occur. Therefore, noncooperation and harassment do not increase the market power of insiders; the equilibrium wage is equal to R_I and only voluntary unemployment may exist. The effective equilibrium work force of each firm λ_1 is given by the equation $Af'(A\overline{m}) = Af'(\lambda_1) = 1$.

In case of an intermediate number $(\underline{m} < m < \overline{m})$ of insiders (scenario 2) their marginal product is higher than their reservation wage: $Af'(Am) > R_I$ at m. With full cooperation and no harassment, outsiders would be profitable. In scenario 2 withdrawal of cooperation and harassment activities ensure that entrants are never profitable, that is, their reservation wage R_E $(> R_I = 1$ because of harassment) is above their marginal product $a_E f'(Am) = 1f'(Am)$. Since insiders are able to make entrants unprofitable, individualistic wage setting by insiders will force the firm to pay

$$(1) \quad R_I < W_I = Af'(\lambda) = Af'(Am);$$

that is, the profit of the marginal incumbent worker is zero.

It is clear that a wage above Af' would induce the firm to fire the insider. Lindbeck and Snower denote this no-firing condition $(W_I \leq Af')$ as absolute profitability con-

*Department of Economics, University of Technology, Vienna, Argentinierstrasse 8/175, Vienna, Austria.

[1] The same notation as in Lindbeck and Snower is used. λ represents effective work force and Ω stands for the utility of a worker. $f(\lambda) \equiv f(a_I L_I + a_E L_E)$ denotes output, a_I is the level of cooperation among insiders, L_I (L_E) is the number of employed insiders (entrants), a_E stands for the level of cooperation between insiders and entrants, h_E is the harassment activity of an insider, H_E denotes the aggregate level of harassment against an individual entrant. There are upper and lower bounds on H_E and a_I (a_E): $0 \leq H_E \leq H$, $1 \leq a_I$, $a_E \leq A$. $R_I(R_E)$ stand for reservation wages of insiders (entrants) and R_I equals 1 whereas $R_E = 1 + H_E \leq 1 + H$. W_I (W_E) denotes insider (entrant) wages. Throughout this comment it is assumed that the potential work force in efficiency units $\bar{s}$ is higher than labor demand; that is, there is always some voluntary or involuntary unemployment.

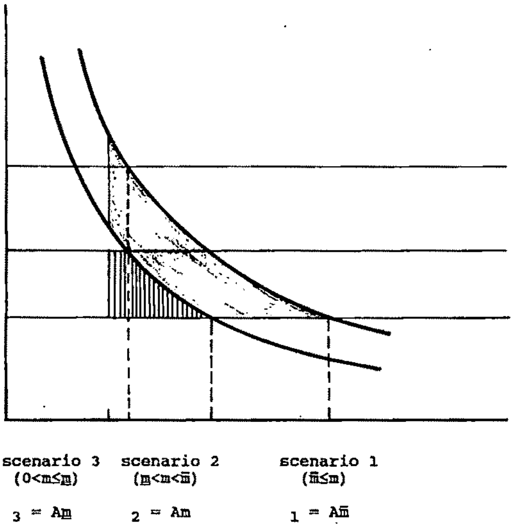


FIGURE 1

straint (APC). In this scenario the effective equilibrium work force is given by $\lambda_2 = Am$, for example $\lambda_2 = A\tilde{m}$ in Figure 1.

If the number of insiders is very low ($m \leq \underline{m}$, scenario 3) the marginal product of entrants is high enough to render them profitable despite withdrawal of cooperation and harassment activities. The marginal insider must make himself at least as profitable as the marginal entrant; otherwise he will be replaced. Lindbeck and Snower denote this requirement the relative profitability constraint (RPC). The insider wage according to the RPC is lower than Af' but higher than R_I:

$$(2)\quad R_I < W_I = A(1+H_E) \leq Af'(Am).$$

It is obvious that the higher H_E the higher W_I. Therefore, $H_E = H$. Entrants receive a compensation equal to

$$(3)\qquad W_E = R_E = R_I + H_E = 1 + H.$$

Equation (3) implies that outsiders are hired up to the point where $f'(\lambda) = f'(\lambda_3) = 1 + H$. Thus, $\lambda_3 = A\underline{m}$. In case of $m < \underline{m}$ the number of entrants employed is given as $A\underline{m} - Am = a_E L_E = L_E$ because insiders do not cooperate with entrants. In scenarios 2 and 3 unemployment is involuntary because if outsiders received full cooperation ($a_E = a_I = A$) and no harassment occurred (i.e., identical conditions of employment, ICE), they would be willing to work for less than the prevailing wage:

$$(4a)\quad R_E(H_E = 0) = 1 < W_I \quad \text{(scenario 2)}$$

$$(4b)\quad R_E(H_E = 0) = 1 < W_E \quad \text{(scenario 3)}$$

I. The Total Profitability Constraint (TPC)

The TPC requires that it is not profitable to replace *all* insiders. Lindbeck and Snower deal with this possibility in fns. 12, 15, and 17. Henceforth I call this scenario zero. They assume in fn. 17 that the upper bound on H_E is itself constrained to be below H^C: $0 \leq H_E \leq H \leq H^C$. The value of H^C is given by the requirement that the maximized profit in scenario 3, π_3^*, is equal to the maximized profit in scenario 0, π_0^*. From the fact that $\delta\pi_3^*/\delta H_E < 0$ and $\pi_3^*(H^C) = \pi_0^*$ they derive the "no-replacement condition" $H_E \leq H^C$. In this section it will be shown that $\pi_3^* = \pi_0^*$ only holds if $H_E = H^C = 0$; that is, in the case of $H_E > 0$ it is profitable to replace *all* insiders.

According to the assumptions of Lindbeck and Snower, entrants are unable to raise their productivity through cooperation and always receive R_E. Therefore, if only entrants are employed, $a_E = 1$, $H_E = 0$, $W_E = R_E = 1$. Profits are maximized at $\lambda_0 = g(1)$ with $g \equiv (f')^{-1}$. From this, π_0^* is given by

$$(5)\qquad \pi_0^* = f[g(1)] - g(1).$$

Notice that λ_0 is in the range of an intermediate incumbent work force: $A\underline{m} < \lambda_0 < A\bar{m}$ (see Figure 1). From (2) and (3), it follows that $\lambda_3 = A\underline{m} = g(1+H)$ and $L_E = A\underline{m} - Am$. Therefore, π_3^* can be written as

$$\begin{aligned}(6)\quad \pi_3^* &= f(A\underline{m}) - W_I m - W_E L_E \\ &= f[g(1+H)] - A(1+H)m \\ &\quad - (1+H)(A\underline{m} - Am) \\ &= f(A\underline{m}) - (1+H)A\underline{m} \\ &= f[g(1+H)] - (1+H)g(1+H).\end{aligned}$$

Differentiation of π_3^* with respect to H shows that maximized profit is a decreasing function of H:

$$(7)\quad \delta\pi_3^*/\delta H = g'\cdot[f'-(1+H)] - g(1+H) = -g(1+H) < 0.$$

From (5), (6), and (7), it follows that $\pi_0^* > \pi_3^*$ if $H > 0$. Therefore, in scenario 3 all insiders will be replaced if they set their wages according to the rule derived by Lindbeck and Snower and if $H_E = H > 0$. Or put differently: The assumption of $H_E = H = H^C > 0$ (together with monopolistic wage setting by insiders) is not compatible with the assumption of $\pi_3^* \geq \pi_0^*$.

The same conclusion emerges for scenario 2 if the incumbent work force is small enough to ensure $Am < \lambda_0 = g(1)$. Since π_2^* is given by

$$(8)\quad \pi_2^* = f(Am) - Af'(Am)m = f(\lambda_2) - f'(\lambda_2)\lambda_2,$$

it is increasing in λ_2. If $\lambda_2 = g(1)$, $\pi_2^* = \pi_0^*$, but for $\lambda_2 < g(1)$, π_2^* is lower than π_0^*.

One way out of this dilemma would be the formulation of a different wage determination process. Suppose for example that the wage is determined in a Rubinstein-bargaining process (Rubinstein, 1982) between *individual* workers and the firm. The outcome of such a process can be approximated by the asymmetric Nash-Solution if the time period between offers and counteroffers becomes small. (See Binmore 1987; Binmore, Rubinstein, and Wolinski, 1986; Hoel, 1986.)

In scenario 3, $f'(Am) \geq 1 + H = W_E$; that is, in general it is profitable to employ entrants. Therefore, in equilibrium (with $L_I > 0$) the marginal product of insiders is equal to $Af'(A\underline{m})$. According to the asymmetric Nash-Bargaining Solution, the division of this cake results from

$$(9)\quad W_I^{\alpha}\cdot[Af'(A\underline{m}) - W_I]^{1-\alpha} \rightarrow \max W_I \quad \text{s.t. } W_I \geq R_I = 1.$$

Insiders want to maximize W_I whereas employers maximize the marginal profit of an insider, but both are constrained by their respective bargaining powers [$0 < \alpha < 1$ resp. $(1-\alpha)$].[2] The outside-option principle (Sutton, 1986) requires that the reservation wage R_I is not considered as status quo-point but as a lower bound on W_I.

According to (9) W_I is given by

$$(10)\quad W_I = \begin{cases} 1 \text{ if } \alpha Af'(A\underline{m}) < 1 \\ \alpha Af'(A\underline{m}) \text{ otherwise.} \end{cases}$$

It is obvious that the RPC is met because $\alpha Af'(A\underline{m}) < Af'(A\underline{m}) = A(1+H)$. If the TPC is also met, Am insiders and $(A\underline{m} - Am) = L_E$ entrants are employed.

Applying the asymmetric Nash-Bargaining Solution to scenario 2, $(m > \underline{m})$ generates

$$(10')\quad W_I = \begin{cases} 1 \text{ if } \alpha Af'(Am) < 1 \\ \alpha Af'(Am) \text{ otherwise.} \end{cases}$$

The APC is always met; that is, W_I is below $Af'(Am)$ and if the TPC holds, all insiders but no entrants are employed.

If insiders differ with respect to their bargaining power α, (10) or (10′) determines a different wage for each of them. For simplicity I assume now that α is the same for all insiders. It is, however, important to keep in mind that there is no coordination among them and that each bargains individually with his employer. This bargaining structure which underlies the analysis of Lindbeck and Snower may lead to a violation of the TPC. In scenario 3 the TPC requires

$$f(A\underline{m}) - W_I m - (1+H)(A\underline{m} - Am) \geq \pi_0^*.$$

Using the third line of (6), (10) and $f'(A\underline{m}) = 1 + H$ yields

$$(11)\quad (1-\alpha)Af'(A\underline{m})m \geq (\pi_0^* - \pi_3^*).$$

[2] The value of α (resp. $(1-\alpha)$) may be derived from the time preferences of the bargaining parties or from asymmetries in the bargaining procedure. (See Binmore, 1987 and Binmore, Rubinstein, Wolinski, 1986 or Hoel, 1986.)

If (11) is violated the firm replaces all insiders. Since $(\pi_0^* - \pi_3^*)$ is positive it is easy to see that for sufficiently large values of α (11) does not hold. The same argument applies to scenario 2 if $Am < \lambda_0$.[3]

II. Disutility of Harassment Activities

According to Lindbeck and Snower, "it is usually safe to assume that harassment activities are disagreeable to the harassers" (p. 171, fn. 7). Nevertheless they assume that harassment does not affect the well-being of insiders because this assumption "has self-evident implications" for their results. I think it would have been better to describe these self-evident consequences in detail in order to enable the reader to make a better judgment of the relevance of their results.

If harassment vis-à-vis entrants decreases the utility of insiders $(\delta\Omega^i/\delta h_E^i < 0)$ employment and wage decisions change for two reasons: (i) Since the contribution of each harasser (h_E^i) to the aggregate level of harassment (H_E) is likely to be small, it is in general not utility maximizing to choose the upper bound value $h_{E,\max}$. Even $h_E^i = 0$ may be an individually rational outcome. (ii) If workers cannot commit themselves to $h_E^i > 0$, they will not threaten to harass entrants because they (and their employers) know that it would not be in their interest to carry out the threat.

Suppose that $\Omega^i = \Omega^i(W_I^i, h_E^i)$ is the utility function of an insider and that $\Omega_1^i = \delta\Omega^i/\delta W_I^i > 0$, $\Omega_2^i = \delta\Omega^i/\delta h_E^i < 0$.[4] The aggregate level of harassment H_E is given as $H_E = \Sigma_{i=1}^m h_E^i$.

Assuming that the wage is determined by (10) or (10′) and that $\alpha^i = \alpha$, all i W_I^i can be written as $W_I = \alpha A f'(Am + L_E)$.[5] The level of L_E is derived from $f'(Am + L_E) = 1 + H_E$. Therefore, L_E is a function of H_E and $\delta L_E/\delta H_E = (1/f'')$. From the maximization of $\Omega^i[\alpha A f'(Am + L_E(H_E)), h_E^i]$ with respect to h_E^i follows

$$\Omega_1^i \alpha A + \Omega_2^i = 0, \tag{12}$$

in case of an interior solution for h_E^i. It is obvious that the assumption of $\Omega_2^i = 0$ implies $h_E^i = h_{E,\max}^i$. In general, however, $h_E^i < h_{E,\max}^i$ holds and if Ω_2^i is large enough, even $h_E^i = 0$ may be an individually rational outcome.

If we allow for binding threats, (12) determines the optimal threat level of h_E^i. But if we stick to Lindbeck and Snower's assumption of a noncooperative game, we cannot allow for such commitment possibilities. In the absence of self-binding threats, $h_E^i > 0$ is not credible because it can only be implemented after the firm has chosen to employ some entrants.

Suppose that the employer has chosen $L_E > 0$ such that $\lambda = \lambda_0$.[6] This decision *fixes* the wage at $\alpha A f'(\lambda_0)$. *Given* this level of λ, it is never optimal for an insider to execute his threat because $h_E^i > 0$ does not change W_I but reduces Ω^i. Since a rational employer recognizes the incredibility of $h_E^i > 0$, he will always choose L_E such that $\lambda = \lambda_0$ and since a rational worker can foresee this, he will not threaten $h_E^i > 0$.

III. Pareto Improving Contracts

Our argument in Section II implies that if harassment is costly for the insiders, unemployment caused by harassment will vanish. In this section we show that even if harassment is *not* costly for insiders the firm can always "persuade" insiders to refrain from $h_E > 0$. Moreover the same type of contract that does away (in equilibrium) with $h_E^i > 0$ is also capable of inducing insiders to cooperate fully with entrants.

Let us now consider these contracts in more detail: (i) The first contract *guarantees*

[3] In scenario 2 the TPC requires $f(Am) - \alpha A f'(Am)m \geq \pi_0^*$. Adding $Af'(Am)m$ on both sides and using (8), results in $(1-\alpha)Af'(Am)m \geq (\pi_0^* - \pi_2^*)$. $(\pi_0^* - \pi_2^*)$ is positive if $Am < \lambda_0$.

[4] In order to simplify the presentation, Ω_2^i is not dependent on the number of entrants. It seems plausible that $(-\Omega_2^i)$ increases with L_E. This would strengthen the argument below.

[5] For the rest of this comment it is assumed that the TPC is met and that $W_I = \alpha A f'(\lambda)$ exceeds R_I.

[6] This argument holds for any level of $L_E > 0$ such that $\lambda \leq \lambda_0$.

insiders (in a legally enforceable way) a slightly higher *time-rate* wage than they would have received through individual bargaining and demands that they cooperate fully with and do not harass entrants. (ii) The second contract assumes that the firm can observe cooperation and harassment with some positive probability τ. It makes the payment above the level of wages that can be achieved through individual bargaining (i.e., the wage premium) contingent on the behavior of insiders; that is, if the firm discovers noncooperation or harassment, the wage premium is not paid.

We analyze first contract (i). Suppose that without this contract a scenario 3 equilibrium prevails,[7] that the number of insiders is $m' < \underline{m}$, and that each insider gets $\alpha Af'(A\underline{m}) > R_I$. In the case of $H_E = 0$, each firm would employ $(\lambda_0 - Am')$ entrants instead of $(A\underline{m} - Am')$. This would generate additional profits of π_h. In Figure 1 π_h is given by the area $(abcd)$. Applying contract (i) the firm guarantees insiders a wage of $W_I^* = [\alpha Af'(A\underline{m}) + w_h]$ where the premium w_h satisfies $0 < w_h \cdot m' < \pi_h$. In exchange for w_h it demands $h_E^i = 0$. Under this contract each insider and the firm is obviously better off.

Since W_I^* is legally enforceable, the firm cannot deviate from this contract and for the same reason insiders have no incentive to deviate. In the Lindbeck/Snower model (with wage bargaining) insiders want to raise their marginal product as much as possible because their wage depends on $A \cdot f'$. The lower the number of entrants, the higher is $\alpha Af'$. But with contract (i), wages are in some sense independent from Af'. Although the actual marginal product at λ_0 is given by $Af'(\lambda_0)$, insiders can be sure of getting W_I^*. They have no reason to restrict L_E through $h_E^i > 0$. It is important to notice that the feasibility of this contract does not depend on the observability of harassment by the firm. Since the firm knows that $\delta\Omega^i/\delta h_E^i = 0$, it also knows that insiders have no incentive for harassment.

Deriving the implications of contract (i) for the cooperation behavior of insiders is now straightforward. The firm could make additional profits π_C (compared with a situation where $a_E = 1$, $\lambda = \lambda_0$, and $H_E = 0$) if each insider cooperates fully, that is, chooses $a_E^i = a_{E,\max}^i$.[8] In Figure 1 π_C is given by the area $(a'b'cbc')$.

It guarantees to pay $[\alpha Af'(A\underline{m}) + w_h + w_C]$, where $w_C > 0$, $w \cdot m' \equiv (w_h + w_C) \cdot m' < \pi_h + \pi_C \equiv \delta\pi$ and demands $a_E^i = a_{E,\max}^i$. As before the firm cannot deviate from this contract and insiders have no incentive to do so. The resulting level of the equilibrium work force is then $A\bar{m}$.

One might argue that although contract (i) gives insiders no incentive to deviate from $a_{E,\max}^i$ and $h_E^i = 0$, it provides also no incentives to stick to these levels. The resulting equilibrium will, therefore, be weak.[9] This problem can be overcome if there exists a positive probability τ that an insider who chooses $h_E^i > 0$ or $a_E^i < a_{E,\max}^i$ loses w. And that is exactly what can be achieved through contract (ii).[10] Starting from the same scenario 3 equilibrium as above, the firm offers a wage $W_I^* = \alpha Af'(A\underline{m}) + w$, $0 < wm' < \delta\pi$.

But now insider behavior is observed at no cost to the firm with probability $\tau > 0$[11] and,

[7] The following argument applies also to scenario 2 as long as $Am < \lambda_0$.

[8] $a_{E,\max}^i$ is the maximum level of cooperation of an individual insider toward entrants. Of course if $a_E^i = a_{E,\max}^i$, all i, $a_E = A$ follows.

[9] According to Harsanyi (1986, p. 104), an equilibrium is called strong if each player's (i.e., insider's and firm's) equilibrium strategy is the *only* best reply to the other players' equilibrium strategies. Otherwise it is called weak. Under contract (*i*) any level of a_E^i and h_E^i with $0 \le a_E^i \le a_{E,\max}^i$, $0 \le h_E^i \le h_{E,\max}^i$ is a best reply; hence the equilibrium is weak.

[10] Of course, if a_E^i and h_E^i are *perfectly* observable *and* verifiable in court and the firm can make the payment of w contingent on $a_E^i = a_{E,\max}^i$, $h_E^i = 0$. But contract (*i*) assumes neither perfect observability nor verifiability in court.

[11] The employer's getting this information costlessly (with some minimum probability) does not seem to be a strong assumption. In most (hierarchical) production processes a minimum of vertical supervision arises as a joint product. Probably there is some degree of horizontal supervision because entrants obviously dislike harassment and prefer to work in a cooperative environment. If they are harassed or do not receive full cooperation, they may complain about it. All that is needed for the validity of the argument below is $\tau > 0$ and $w > 0$. τ and w may be very low.

therefore, if $a_E^i < a_{E,\max}^i$ or $h_E^i > 0$, the expected income of an insider is given by $\{\tau\alpha Af'(A\underline{m}) + (1-\tau)[\alpha Af'(A\underline{m}) + w]\} = \alpha Af'(A\underline{m}) + (1-\tau)w$. In the case of full cooperation and no harassment, income is $\alpha Af'(A\underline{m}) + w$. It is obvious that each insider prefers a no-harassment/full cooperation strategy under this contract. Since $w > 0$ they accept such a contract and since $\delta\pi - wm' > 0$ firms will offer them.

There exists an argument against the implementability of contract (ii), namely, that firms may cheat insiders and claim falsely that $a_E^i < a_{E,\max}^i$ or $h_E^i > 0$ has been observed. In order to remove this incentive to cheat, the contract may stipulate that any premium which is not paid because of noncooperation or harassment has to be distributed to other insiders.

Contract (i) and contract (ii) do away with noncooperation and harassment; that is, in equilibrium there is no incentive for $a_E^i < a_{E,\max}^i$ or $h_E^i > 0$. This induces firms to choose the maximum level of λ that is compatible with their production possibility frontier and the disutility of work (without harassment): $Af'(\lambda) = 1$ of $\lambda = A\bar{m}$. Although in this model we can never get more employment in efficiency units than $A\bar{m}$ (in this sense $\lambda = A\bar{m}$ is a full employment equilibrium) and although in a contract (i)—or contract (ii)—equilibrium any outsider can get a job at his reservation wage, we may have involuntary unemployment if we apply the definition of Lindbeck and Snower. According to their definition, a worker without a job is involuntarily unemployed if his reservation wage in efficiency units and under identical conditions of employment (R_E^{ICE}/a_E^{ICE}) is strictly below the efficiency wage of *insiders* (W_I/a_I). Assuming for simplicity that $a_E^{ICE} = A$, we have $R_E^{ICE}/a_E^{ICE} = 1/A$. In our contract (i)—or contract (ii)—equilibrium W_I is *always* above 1 because insiders can always enforce at least a wage of $R_I = 1$ through individual bargaining and get in addition $w > 0$. Therefore, we have $W_I/a_I = W_I/A > 1/A$, and hence any worker without a job is involuntarily unemployed.[12] In our view it is more appropriate to compare $R_E^{ICE}/a_E^{ICE} = 1/A$ with the efficiency wage of *entrants*, W_E/a_E. Without harassment entrants get $W_E = 1$ and with full cooperation $a_E = A$ (see fn. 8). Thus, implementing contract (i) or (ii) leads to an equilibrium with no involuntary unemployment.[13]

[12] If we assume $a_E^{ICE} < A$, we may still have $W_I/A > 1/a_E^{ICE}$.

[13] We want to stress that our arguments in Section III are also applicable to union models of unemployment (for example, McDonald and Solow, 1981; Johnson and Layard, section 5.1, 1986). These models are implicitly or explicitly based on the *assumption* that all workers in a firm have to be paid the same nonmarket clearing wage. It is, however, not clear why insiders (union workers) should object to the employment of some outsiders at a market clearing wage if they are guaranteed their jobs and get a wage premium which gives them a higher total income than they would have received through collective bargaining.

REFERENCES

Binmore, Ken, "Nash Bargaining Theory II," in Ken Binmore and Partha Dasgupta, *The Economics of Bargaining*, Oxford: Basil Blackwell, 1977.

______, **Rubinstein, Ariel and Wolinsky, Asher,** "The Nash Bargaining Solution in Economic Modelling," *Rand Journal of Economics*, Summer 1986, *2*, 176–88.

Fehr, Ernst, "Do Cooperation and Harassment Explain Involuntary Unemployment?" Arbeitspapier 1988-7 des Arbeitskreises Sozialwissenschaftliche Arbeitsmarktforschung.

Harsanyi, John C., *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge: Cambridge University Press, 1986.

Hoel, Michael, "Perfect Equilibria in Sequential Bargaining Games with Nonlinear Utility Functions," *Scandinavian Journal of Economics*, June 1986, *2*, 383–400.

Johnson, G. E. and Layard, P. R. G., "The Natural Rate of Unemployment: Explanation and Policy," in O. Ashenfelter and R. Layard eds., *Handbook of Labor Economics*, Amsterdam: North-Holland, 1986.

Lindbeck, Assar and Snower, Denis, "Cooperation, Harassment, and Involuntary Unemployment: An Insider-Outsider Approach," *American Economic Review*, March 1988, *1*, 167–89.

——— and ———, "Explanations of Unemployment," *Oxford Review of Economic Policy*, June 1985, *2*, 34–59.

McDonald, Ian M. and Solow, Robert M., "Wage Bargaining and Employment," *American Economic Review*, December 1981, *5*, 896–908.

Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, January 1982, *1*, 97–109.

Sutton, John, "Noncooperative Bargaining Theory; An Introduction," *Review of Economic Studies*, October 1986, *5*, 709–24.

Cooperation, Harassment, and Involuntary Unemployment: Reply

By ASSAR LINDBECK AND DENNIS J. SNOWER*

The basic idea of our (1988) article is that involuntary unemployment can occur because (a) the experienced, incumbent employees, "insiders," have an incentive to keep their wages above the market-clearing level and (b) these insiders are able to achieve this (within limits) by imposing their own preferences on their firms and on the other workers, "outsiders," through the threat of harassing and withdrawing cooperation from underbidders. The outsiders are involuntarily unemployed in the sense that their inability to find jobs is due to the circumstance that they face less favorable working conditions (in terms of cooperation and harassment) than insiders do.

The involuntary unemployment persists because the discriminatory conditions facing the outsiders are not eliminated through underbidding. In particular, the outsiders may be disinclined to engage in underbidding because they know that if they did so, the insiders would "harass" them, thereby raising their reservation wage. Moreover, firms may be unwilling to replace some of their insiders by underbidding outsiders, because they know that if they did so, the remaining insiders would not "cooperate" with the underbidders in the production process, thereby reducing the underbidders' productivity.

It is also worth noting that firms may be disinclined to replace *all* their insiders by outsiders. In our (1988) article we simply took this for granted, but Ernst Fehr's comment points to the need to provide specific reasons why this may be so. As we will show, these are not hard to come by.

For clarity and brevity, we used a particularly simple model to show how cooperation and harassment activities can give rise to involuntary unemployment. Our model is appropriate to the problems we chose to analyze; clearly, if the domain of problems is broadened, the model must be extended. The simplicity of our analytical framework inevitably leaves some potentially interesting questions unanswered, and Fehr's comment raises three of them. Let us consider each in turn.

I. Replacement of Insiders

If insiders protect their positions by withholding cooperation from underbidders and harassing them, then why does the firm not replace all its insiders by outsiders?

In practice, firms rarely replace their entire insider work forces, presumably because they expect this action to be extremely costly. A potentially important reason for this is that (i) firms face labor turnover costs, which means that employment activities are associated with economic rent and (ii) the insider wage is the outcome of negotiations in which insiders capture only some of this rent. Then the wage cost of a firm's insider work force exceeds that of an entrant work force by no more than the labor turnover costs associated with the replacement of the insider work force. By implication, the firm would not find it profitable to implement the "replacement strategy," whereby all insiders are replaced by new entrants.

The firm's turnover costs from the replacement strategy may take many forms, but the following are particularly relevant to our (1988) article:

- The new work force may be less productive than the one it replaces, because

*Institute for International Economic Studies, University of Stockholm, S106 91, Stockholm, Sweden, and Department of Economics, Birkbeck College, University of London, 7-15 Gresse Street, London W1P 1PA, England, respectively.

newcomers generally have less on-the-job training; for example, they usually have not had the opportunity to acquire the skills required to cooperate with one another in team production.

— Cooperation and on-the-job training are often acquired by learning from the existing insiders; but if all insiders have been fired, the learning process may well be lengthy, uncertain, and expensive.

— The insiders' harassment activities may be expected to continue even if all insiders have been fired, as fired employees are generally capable of performing such activities outside the firm. These activities may take the form of picket lines and social ostracism, as well as other forms of industrial and social unrest, perhaps even sabotage and violence. They drive the reservation wages of the underbidding entrants above those that the insiders had when they were employed.

— There may be a loss of customer goodwill (associated with a loss of firm revenue), following the adverse publicity which the replacement of entire work forces is likely to attract.

In addition, of course, firms often face a wide variety of labor turnover costs stemming from job security legislation and union activities (such as strikes and work-to-rule actions).

In our (1988) article we attached little importance to these considerations, since the possibility of replacing entire insider work forces—especially under conditions of team production—struck us as remote. Rather, we implicitly assumed that impediments to such replacement exist, and we concentrated on what seemed to be a practically more essential question, namely, whether firms may be unwilling to replace even *some* of their insiders by underbidding entrants due to the insiders' cooperation and harassment activities.

Fehr's comment points to the need for giving the replacement strategy some formal attention. Our model shows that, under individualistic bargaining, the insiders seek to achieve the highest possible wage subject to the constraint that no insider becomes unprofitable to the firm (the "absolute profitability constraint") and that each insider is at least as profitable as the marginal entrant (the "relative profitability constraint"). Fehr suggests a third constraint, namely, that the entire insider work force remain at least as profitable as a work force of entrants (the "total profitability constraint").

Fehr's analysis implies that the total profitability constraint on the insider wage will be binding if *all* of the following conditions are met: (i) fired incumbents do not harass the entrants who replace them, (ii) the firms' incumbent work forces are sufficiently small, (iii) the insider wage is set unilaterally by the insiders, (iv) entrants' productivity is independent of the number of insiders in the firm, and (v) there are no rent-related turnover costs other than those associated with the insiders' cooperation and harassment activities. Although this result is correct, it is quite unlikely for all these conditions to be met simultaneously in practice. To incorporate the replacement strategy into our analysis, these conditions clearly need to be relaxed.

First, as noted, the mere fact that all the insiders have been fired does not preclude them from harassing the workers who have ousted them. Accordingly, let us suppose that the new entrants will be harassed by the fired insiders,[1] so that the entrant wage is $W_E = (1+H)$ (where H is the disutility from being harassed and the reservation wage in the absence of harassment would be $W_E = 1$). Then the firm's profit under the replacement strategy is

$$(1) \qquad \pi_R^* = f[g(1+H)] - (1+H)\cdot g(1+H).$$

This is equal to the firm's profit (π_3*) under Scenario III (in which all insiders are retained and some entrants are hired) and is

[1] We did not make this assumption in our (1988) article, as the replacement strategy plays no role in our formal analysis. In making this assumption now, we show that the replacement strategy is unprofitable, while our formal analysis remains unaffected.

less than its profit (π_2^*) under Scenario II (in which the insiders are retained but no entrants are hired), as given by equations (6) and (8) of Fehr's comment. Thus, under the assumption that fired workers are able to engage in harassment activities, the firm has no incentive to implement the replacement strategy.

Second, when considering the potential profitability of replacement, the assumption concerning the division of economic rent becomes significant. In this context, it is desirable—as Fehr suggests—to portray the wage as the outcome of a bargaining process between the firm and its insiders, whereby the insiders capture only a fraction α ($0 < \alpha < 1$) of the available rent. In that case, the firm's profit under the replacement strategy (π_R^*), given by equation (1) above, is always strictly less than π_2^* and π_3^*, and thus the insider work force will not be replaced.

Finally, in evaluating the replacement strategy it is important to observe that the productivity of entrants generally depends on the number of insiders cooperating with them and training them. As more insiders are replaced by entrants, the entrants lose access to colleagues who can cooperate with them and provide on-the-job training, and thus the less productive the entrants may become. It was unnecessary to incorporate this consideration into our model when we only analyzed the replacement of marginal insiders; but the possibility of replacing the entire insider work force makes the matter significant.

To formalize this dependence of entrant productivity on insider availability, let L_I and L_E be the number of insiders and entrants in the firm (respectively), a_I and a_E be their respective labor endowments, and a_I^i be the degree to which the representative insider (i) cooperates with each entrant. Then the firm's production function may be respecified[2] as

$$(2) \qquad Q = f(a_I \cdot L_I + a_E \cdot L_E),$$

where

$$a_E = \mu(a_E^i, L_I), \qquad \mu_1, \mu_2 > 0,$$

and

$$\mu[\min(a_E^i), 0] < 1.$$

Then the firm's employment level in Scenario III is

$$(3) \quad \underline{\lambda} = g[(1+H)/\mu[\min(a_E^i), m]]$$

and its profit is

$$(3a) \qquad \pi_3^* = f[\underline{\lambda}] - (1+H) \cdot \underline{\lambda} > 0.$$

Yet its profit under the replacement strategy is

$$(3b) \qquad \pi_R^* = f[\mu[\min(a_E^i), 0] \cdot L_E^*] - (1+H) \cdot L_E^*.$$

Since $\mu[\min(a_E^i), 0] < 1$ it is easy to show that $\pi_R^* < \pi_3^*$; in other words, it is less profitable to replace all the insiders than to retain them.

In short, there are a host of good reasons why firms do not fire *all* their insiders if these try to prevent underbidding through noncooperation and harassment activities. Given any of these reasons, the total profitability constraint becomes redundant, and the firm has no incentive to replace the entire insider work force.

II. Credibility

If insiders find it disagreeable to engage in harassment, is their threat to harass potential entrants credible?

In our (1988) model, insiders' harassment decisions cannot be implemented until the firm's employment decisions have been made, and the employment decisions in turn are made only after the insider wage has been set. This is the background to Fehr's point that insiders have nothing to gain from harassing entrants—their wage and employ-

[2] In our (1988) article, we assumed that $a_E = 1$ when $a_E^i = \min(a_E^i)$ and when all insiders are retained. We now assume that $a_E < 1$ when $a_E^i = \min(a_E^i)$ and all insiders are fired.

ment has already been determined.[3] But if the harassing activity is disagreeable to them, then insiders do have something to lose. In that event, the harassment threat is not credible: the insiders have no incentive to implement their harassment threat, and consequently the firm will ignore this threat when making its employment decisions.

This argument is merely an artifact of combining the assumption on the disagreeableness of harassment with the one-period framework of analysis. For brevity and simplicity, our (1988) model assumed that agents have a one-period time horizon, and we sidestepped the credibility problem by supposing that insiders suffer no utility loss from harassing entrants. Yet if the credibility problem is to be examined in detail, this approach needs to be amended.

To begin with, it is important to note that although harassing may be an inherently disagreeable activity, the insiders may in fact derive positive utility from it when they are provoked by entrants engaging in underbidding.[4] (Consider, for example, common attitudes of incumbent workers to so-called "scabs.") In short, whether the insiders gain utility or disutility from harassing may well depend on whether this activity is a response to a threat to insiders' income or job security. In context of our (1988) analysis, harassment may give the insiders utility since it is assumed to occur only when there is underbidding. In that event, the harassment threat is credible in our model, even if the insider wage is predetermined when the insiders make their harassment decision.

Furthermore, even if harassing activities would *always* be disagreeable to the insiders, the harassment threat may nevertheless be credible in a multiperiod setting. In that context, the insiders may have a current incentive to engage in harassment activities in order to establish a reputation in the future. This reputation can discourage the firm from future hiring: thus the firm's work force will be smaller than it would otherwise have been and, by implication, the insider wage will be larger. In particular, the insiders will have an incentive to harass the entrants whenever the disutility from doing so is less than the utility from the associated rise in the *future* insider wage.

To make this point formally, let us assume that workers maximize their utility over an infinite time horizon. As in our (1988) paper, our analysis of insiders' harassment activity applies either (a) to an individual insider, acting atomistically within a production team or (b) to a firm-specific union of insiders that is able to impose its decisions on its members. Let Γ_t be the level of harassment of an entrant (either by an individual insider or by a union of insiders) in time period t, measured by the entrant's disutility from being harassed in that period. For simplicity, but without loss of generality, let the harassment level be a discrete variable, so that $\Gamma_t = H$ (where H is a positive constant) when harassment takes place and $\Gamma_t = 0$ when it does not.[5] Let H^e_{t+1} stand for the firm's expecta-

[3] Fehr also argues that insiders have no incentive to harass entrants when harassing activities yield significant disutility and when each insider makes an insignificant contribution to the overall level of entrant harassment. This argument misses the mark. Our entire analysis is based on the assumption that either (a) the firm's team is sufficiently small for each insider to make a significant contribution to the overall levels of cooperation and harassment or (b) the insiders in each firm belong to a union which is able to impose its cooperation and harassment decisions on its members. Clearly, in the absence of such an assumption, insiders have no individual incentive to engage in discriminatory cooperation and harassment practices.

[4] We failed to clarify this matter in fn. 7 of our article, which however was not concerned with the credibility issue. Yet setting the record straight does not affect the conclusions of our analysis.

[5] It is straightforward, though algebraically more complicated, to derive our qualitative results from the assumption that harassment is a continuous variable. Then, as Fehr indicates, the harassment activity level is set so that the marginal gain from harassment (via the induced rise in the insider wage) is equal to the marginal disutility of the harassment activity per se. Observe that when the production team is reasonably small (as production teams often are), the marginal gain from harassment may be quite large; and when the harassment activity has been provoked through underbidding, the associated marginal disutility may be quite small. In that event, the optimal level of harassment may well be large.

tion of Γ_t, and we assume that this expectation is formed as follows:

$$(4) \qquad H^e_{t+1} = \Gamma_t$$

(which can be shown to be rational in the context of our model).

Thus, when harassment takes place in period t ($\Gamma_t = H$), the firm expects harassment to continue in period $t+1$ ($H^e_{t+1} = H$). Then the insiders receive the wage $W_{I,t+1} = A \cdot (1 + H)$ (where the insider-entrant productivity ratio is A and the entrant's reservation wage in the absence of harassment is 1). On the other hand, if there is no harassment in period t ($\Gamma_t = 0$), the firm expects no harassment in the next period either ($H^e_{t+1} = 0$). Then the insider wage is $W_{I,t+1} = A$.

In line with Fehr's assumption that harassment is disagreeable to the harassers, let the positive constant h_E be the insider's (or union's) level of harassing activity, measured by the disutility of harassing. Clearly, the level of harassment (H) affecting the entrant is positively related to the level the harassing activity (h_E) by the insider (or union); for simplicity, let $H = \beta \cdot h_E$. Let δ be the rate of time discount. Then, assuming that the firm expects harassment in the first period, the present value of the insider's (or union's) utility when there is harassment in all periods is

$$(5a) \qquad PV_h = [A \cdot (1+H) - h_E] \cdot (1 + \delta + \delta^2 + \cdots) = [A \cdot (1 + \beta \cdot h_E) - h_E] \cdot [1/(1-\delta)].$$

When there is no harassment in any period, the present value is

$$(5b) \qquad PV_n = A \cdot (1+H) + A \cdot (\delta + \delta^2 + \cdots) = A \cdot (1 + \beta \cdot h_E) + A \cdot [\delta/(1-\delta)].$$

One of these two present values represents an optimum for the insider (or union) since the present values depend linearly on the per-period utility, and thus the insider can always make himself at least as well off by always harassing or never harassing than by harassing in some periods but not in others.

The harassment threat is credible iff

$$(6a) \qquad PV_h \geq PV_n,$$

which implies that

$$(6b) \qquad h_E \cdot (\beta \cdot \delta - 1) \geq 0.$$

Condition (6b) holds whenever $\beta \geq (1/\delta)$. In that event, insiders' harassment threat is credible even when the harassing activity per se yields disutility.

III. Pareto-improving Contracts

Can Pareto-improving time-rate wage contracts be found which induce insiders to cooperate fully with entrants and to forego harassing them, thereby eliminating the involuntary unemployment?

Fehr suggests two contracts which allegedly perform this function. The first contract grants the insiders a wage premium (w) in addition to what they would receive under individualistic bargaining ($\alpha \cdot f'(A \cdot \underline{m})$,[6] *provided* that they cooperate fully with the entrants and forego harassing them. Fehr assumes that this contract can be implemented even if harassment (and presumably also cooperation) is unobservable to the firm. This means that the firm must offer each insider a wage of $W_I^* = [\alpha \cdot f'(A \cdot \underline{m}) + w]$, *regardless* of the insider's actual cooperation and harassment activities. Then it is easy to see that the insiders will cooperate fully with entrants and avoid harassing them only when they have nothing to gain from doing otherwise. But then the cooperation and harassment threats are not credible. Consequently, the insiders will not make these threats in the first place, and thus the contract is unnecessary.

Now suppose that the insiders *do* have something to gain from implementing these

[6] Recall that α is a measure of insider bargaining power and $f'(A \cdot \underline{m})$ is the marginal product of the firm's work force.

threats—as when they gain utility from harassing or withholding cooperation in response to underbidding, or when they establish a reputation that discourages the firm from future hiring.[7] Then the contract will not be effective in promoting cooperation and discouraging harassment, for precisely the same reason that the regular time-rate wage (in our model) is ineffective.

The second suggested contract looks more promising, at first sight. This contract grants each insider the wage $W_I^* = \alpha \cdot f'(A \cdot \underline{m})$ when the firm observes him to have harassed entrants or withheld cooperation from them, and gives him an additional wage premium otherwise (so that his wage becomes $W_I^* = [\alpha \cdot f'(A \cdot \underline{m}) + w_h]$. The underlying assumption is that the firm can, with some positive probability, observe each insider's cooperation and harassment activities—observe them, in fact, with sufficient accuracy and objectivity so that every insider can agree with the firm about the circumstances under which a wage reduction for noncooperation or harassment is justified.

This assumption is quite unlikely to be met in practice, and this may be a reason why we do not observe such contracts in the real world. Cooperation activities are the product of teamwork among workers, and it is an inherently impossible task to identify an individual's contribution to a cooperative effort with sufficient objectivity and accuracy to use that observation as a basis for wage payments. It is difficult to imagine what procedures a firm could use to convince workers, arbitrators, or courts that a *particular* insider had been uncooperative, rather than the other insiders and entrants of the team. The difficulties involved in observing harassment activities are, if anything, even greater since the final product of these activities (a rise in the entrant's disutility of work) is not open to objective measurement.

In the absence of accurate and objective observations of cooperation and harassment activities by individual insiders, the firm could face substantial litigation costs whenever it refuses to pay the wage premium, given that the above contract is in force. These costs, along with a possible loss of customer goodwill, may generally be expected to deter firms from offering the above contract. Moreover, in the hypothetical event that accurate and objective observation could be made, the above contract may still not keep the insiders from their rent-seeking activities, if (as already discussed before) the insiders derive utility from withholding cooperation or harassing underbidders or if they have the opportunity of building reputations in a multiperiod context.

It is worth noting, however, that whereas the inputs into team production cannot be measured objectively in practice, the output *may* be measurable in some cases. Thus output-related wage contracts appear to be more promising devices for inducing insiders to cooperate with entrants. Yet such contracts also give rise to problems of implementation and our (1988) paper discusses some such difficulties. Finally, observe that, like Fehr's wage contracts, the output-related ones are incapable of reducing involuntary unemployment which arises from insiders' harassment activities.

In conclusion, it appears to be very difficult to design practically implementable contracts that bribe the insiders to forego their discriminatory cooperation and harassment activities, and Fehr has certainly not succeeded in doing so.

[7] If future hiring is discouraged, the insider wage will be greater than it otherwise would have been. Contrary to Fehr's assertion, the insider wage ($W_I^* = [\alpha \cdot f'(A \cdot \underline{m}) + w]$) *does* depend on the marginal product of the firm's work force ($f'(A \cdot \underline{m})$), and thus insiders *do* have an incentive to restrict entry to this work force.

REFERENCE

Lindbeck, Assar and Snower, Dennis J., "Cooperation, Harassment, and Involuntary Unemployment: An Insider-Outsider Approach," *American Economic Review*, March 1988, *78*(1), 167–88.

Internal Migration and Urban Employment: Comment

By Alan Day Haight*

Recently in this journal William E. Cole and Richard D. Sanders (1985) criticized the Todaro migration model and offered a different approach. A lively and interesting debate followed, but the Cole and Sanders (CS) model was not actually solved. Indeed, Michael Todaro (1986, p. 566) suggested that the model yielded no unique algebraic solution, a charge to which CS (1986) did not respond. This note identifies the changes needed to provide closure of the CS model.

Combining all of the [CS] equations on the subsistence sector (Section VI), and inserting a full-employment equation,[1] one obtains,

$$F^{-1}\left[X\left[\hat{W}_{us}N_{us}/F(N_{us});\bar{P}_F,\bar{Y}_m,\overline{Pop}\right]+\bar{G}\right]$$

$$=\bar{N}_s-\frac{\bar{P}_r\bar{Q}_{rs}}{\hat{W}_{us}},$$

where the notation is that of CS. (Bars are added over characters to indicate constants.[2]) Unfortunately, this is one equation in two unknowns: the urban-subsistence wage ($\hat{W}_{us}$) and urban-subsistence employment (N_{us}).

To close the model, I suggest adding an equation requiring that the production function for urban-subsistence services be **linear:** $F(N_{us})=\bar{c}N_{us}$, where $\bar{c}$ represents the average (and marginal) product of urban-subsistence labor. (Evidently CS had both full employment[3] and linear production[4] in mind, but omitted specifying them.) This linearity eliminates the troublesome $N_{us}/F(N_{us})$ factor, leaving

$$\frac{1}{\bar{c}}\left[X\left[\hat{W}_{us}\frac{1}{\bar{c}};\bar{P}_F,\bar{Y}_m,\overline{Pop}\right]+\bar{G}\right]$$

$$=\bar{N}_s-\frac{\bar{P}_r\bar{Q}_{rs}}{\hat{W}_{us}},$$

where the LHS and RHS are the derived demand for urban-subsistence labor and residual supply of urban-subsistence labor, respectively. This reduced form does determine $\hat{W}_{us}$, closing the model.

*Instructor, Department of Economics, Bates College, Lewiston, ME 04240, and doctoral candidate, University of Wisconsin-Madison.

[1] The full employment equation inserted here is $N_{us}+N_{rs}=\bar{N}_s$, where "$\bar{N}_s$" represents total subsistence population, employed in urban-subsistence and rural-subsistence sectors.

[2] To verify which variables are exogenous, see CS (1985, notes 29, 35, 41, 42). CS give no equation endogenizing G (hence $\bar{G}$ is used above), but CS do sketch an example (note 33) where $\hat{G}$ is an export base multiple of $\hat{X}$. If one used a function such as $G[X(W_{us},N_{us})]$ instead of exogenous $\bar{G}$, this would not alter the closure issue: G would then be an indirect function of the same unknowns as X. Calculations are available from the author.

[3] For their position on full employment and lack of barriers to entry, see CS (1985, p. 482, p. 490).

[4] CS note that labor is the only input (equation 7) and assume that the competitive U–S sector pays labor its average product (equation 8).

REFERENCES

Cole, William E. and Sanders, Richard D., "Internal Migration and Urban Employment in the Third World," *American Economic Review*, June 1985, *75*, 481–94.

______ and ______, "Internal Migration and Urban Employment: Reply," *American Economic Review*, June 1986, *76*, 570–72.

Todaro, Michael P., "Internal Migration and Urban Employment: Comment," *American Economic Review*, June 1986, *76*, 566–69.

Cooperative and Noncooperative R&D in Duopoly with Spillovers: Comment

By IRENE HENRIQUES*

Claude d'Aspremont and Alexis Jacquemin (1988) employ a simple yet elegant symmetric duopoly model of R&D and spillovers to compare several equilibrium concepts. These concepts include (1) the two-stage noncooperative solution, (2) the two-stage mixed game,[1] (3) the two-stage fully cooperative solution,[2] and (4) the social planner's optimum.[3]

For each of the cases stated above, they computed the equilibrium levels of output ($Q = q_1 + q_2$) and R&D ($x_1 = x_2 = x$) and the required second-order conditions. They report (i) for large spillovers (i.e., $\beta > 0.5$) $x^{**} > \overset{''}{x} > \hat{x} > x^*$ and $Q^{**} > \hat{Q} > Q^* > \overset{''}{Q}$ and (ii) for small spillovers (i.e., $\beta \leq 0.4$) $x^{**} > \overset{''}{x} \geq x^* > \hat{x}$ and $Q^{**} > Q^* > \hat{Q} > \overset{''}{Q}$, where x denotes a firm's R&D level, Q denotes total industry output, ** denotes the social optimum, $''$ denotes the fully cooperative model, * the noncooperative two-stage case, and $\hat{}$ the mixed game. β is the spillover parameter.

Here we show that comparing the pure cooperative and the pure noncooperative solutions as defined by d'Aspremont and Jacquemin is only meaningful when the noncooperative solution is stable, that is, when spillovers are not too small. We find that, for very small spillovers (in our example this occurs when $\beta \leq 0.17$), the d'Aspremont–Jacquemin observation holds because the noncooperative model is unstable. The importance of this result rests on the fact that even though the output reaction functions cross correctly when $\beta \leq 0.17$, the R&D reaction functions do not. When $0.17 < \beta < 0.41$, stability obtains but R&D levels are higher in the noncooperative case than the fully cooperative one. For large spillovers the d'Aspremont–Jacquemin result is confirmed. Moreover, we find that the introduction of spillovers in the case of the noncooperative model tends to "*promote*" stability. In the case of the cooperative model, however, as the level of spillovers is increased, an equilibrium ceases to exist.

I. Stability and Spillovers

Following d'Aspremont and Jacquemin (1988), we use a linear inverse demand schedule for the "industry"

$$p = a - b(q_1 + q_2) \tag{1}$$

such that $a, b > 0$ and $q_1 + q_2 \leq a/b$. Production costs are given by

$$C_i(q_i, x_i, x_j) = (A - x_i - \beta x_j) q_i, \qquad i, j = 1,2 \quad i \neq j, \tag{2}$$

where $0 < A < a$, $A > x_i + \beta x_j$, $0 \leq \beta < 1$ and β is the spillover parameter.[4] The cost of installing x_i are $(\gamma/2)x_i^2$ $(i = 1,2)$. Symmetry results from assuming that the players have the same costs.

Firm i's profits are given by

$$\Pi_i = [a - b(q_1 + q_2)] q_i - [A - x_i - \beta x_j] q_i - (\gamma/2) x_i^2. \tag{3}$$

[*] Department of Economics, Queen's University, Kingston, Ontario Canada K7L 3N6. The author is indebted to John M. Hartwick and Perry A. Sadorsky for their helpful suggestions. I would also like to thank the FCAR fund of Le Ministère de l'Education du Québec for financial support.

[1] Here firms cooperate in R&D but remain noncooperative in output.

[2] Firms cooperate in both their R&D and output decisions.

[3] A two-stage framework was also employed when computing the social optimum.

[4] When $\beta = 0$, we have the Barbara J. Spencer and James Brander (1983) two-stage duopoly R&D game.

The model described above is employed in two different games. In the first, firms act noncooperatively in both output and R&D. In the second game, firm behavior is modeled as a two-stage game played by two firms who *fully* cooperate in both their output and R&D decisions.[5]

Stability in duopoly involves reaction functions crossing "correctly."[6] In the model described by d'Aspremont and Jacquemin, the stability requirements must apply for both choice variables (namely, output and R&D levels). In other words, in R&D and output games, one must reduce the system to reaction functions in either output or R&D space.

The plausible approach, therefore, is to work back to the first stage (R&D stage) and analyze the reaction functions in R&D space. Before working back to the R&D stage, however, let us first look at the reaction functions in output space (equation (4)), taking R&D levels as given (i.e., $x_i \geq 0$).

$$(4)\qquad q_i = \frac{a - bq_j - A + x_i + \beta x_j}{2b}.$$

These output reaction functions will *always* cross "correctly" in our example because $0 < A < a$, $x_i + \beta x_j \leq A$, $b > 0$ and R&D levels are given.

In R&D space, however, we employ the following reduced form reaction functions:

$$(5)\qquad x_i = \frac{-2(2-\beta)\{(a-A)+(2\beta-1)x_j\}/9b}{\dfrac{2(2-\beta)^2}{9b} - \gamma},$$

where $i, j = 1,2$ and $i \neq j$. The numerator of (5) is the second-order condition in the R&D stage that must be less than 0 for an equilibrium. These reduced form reaction functions cross "correctly" when $|\partial x_i / \partial x_j|$ is less than 1. Differentiating (5) with respect to x_j yields[7]

$$(6)\qquad \frac{-2(2-\beta)(2\beta-1)/9b}{\dfrac{2(2-\beta)^2}{9b} - \gamma}.$$

Plotting the reduced form reaction functions in R&D space gives us a further dimension with which to examine the d'Aspremont and Jacquemin results. Figure 1 depicts six cases. The first three cases represent the small spillover situations, while the last three represent the large spillover cases.[8]

From Figure 1 (cases 1 through 3), we observe that the introduction of spillovers leaves the model unstable for $\beta < 3/2 - \sqrt{7/2} \approx 0.17$ and stable for $0.17 < \beta < 1.0$. The existence of instabilities can lead to corner solutions for which specialization in R&D production is implied.[9] Note, however, that for $\beta > 0.6$ an equilibrium cannot be obtained in the *fully cooperative model*.[10] Furthermore, in the case of the noncooperative two-stage model, we observe that although the second-order conditions are satisfied when $\beta < 0.17$, stability is not assured.[11]

[5] The solutions to these games can be found in d'Aspremont and Jacquemin (1988).

[6] See Jesus K. Seade (1980) for an analysis of stability in one-stage Cournot duopoly games.

[7] The requirement that $|\partial x_i / \partial x_j| < 1$ is exactly that used by Seade (1980). In our example this translates as follows:

$$\left| \frac{\dfrac{\partial^2 g_i}{\partial x_i \partial x_j}}{\dfrac{\partial^2 g_i}{(\partial x_i)^2}} \right| < 1 \qquad i, j = 1,2 \text{ and } i \neq j,$$

where $g_i(\cdot) = [a - b\{q_1(x_1, x_2) + (q_2(x_1, x_2)\}] q_i(x_1, x_2) - [A - x_i - \beta x_j] q_i(x_1, x_2) - (\gamma/2)x_i^2 \quad i, j = 1,2 \ i \neq j$.

[8] These reduced reaction functions are, in general, given by $x_1 = x_1(x_2, q_1(x_1, x_2), q_2(x_1, x_2))$ and $x_2 = x_2(x_1, q_1(x_1, x_2), q_2(x_1, x_2))$.

[9] See John M. Hartwick (1988) for further details.

[10] This point is important since it imposes restrictions on the values of β, γ, A, a, and b that one can employ to satisfy all the equilibrium concepts discussed in d'Aspremont and Jacquemin.

[11] In the fully cooperative case, the second-order conditions, $(1+\beta)^2/2 - 2\gamma < 0$, are satisfied for all $\beta \leq 0.6$.

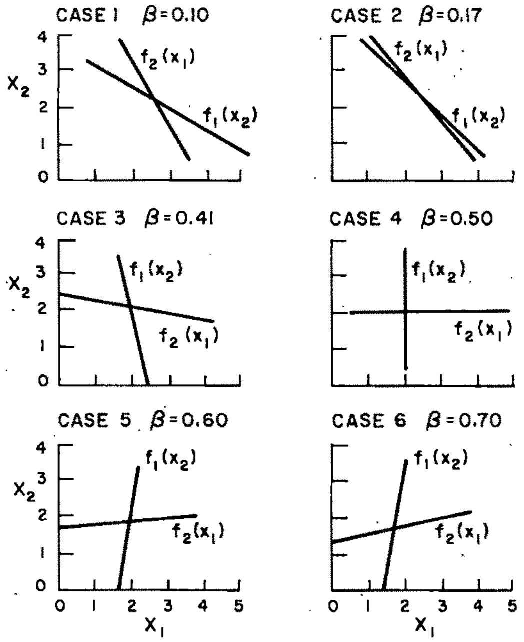


FIGURE 1. REDUCED FORM REACTION FUNCTIONS IN R&D SPACE

ASSUMPTION 1: *Inverse demand*: $p = a - b(\cdot)$, $a = 10.0, b = 1.0$
Production costs: $(A - x_i - \beta x_j)q_i$, $A = 7.0$, $\beta \leq 0.41$
R&D costs: $(\gamma/2)x_i^2, \gamma = 1.0 \quad i, j = 1,2 \quad i \neq j$

Hence it is necessary to set the proper parameter restrictions according to the relevant stability and existence requirements of each model before one can compare both models in either a qualitative or quantitative manner.

From cases 4 through 6, we observe that under large spillovers the slopes of the reduced form reaction functions change. Although these reaction functions cross "correctly," R&D levels (x_1 and x_2) are positively related under large spillovers rather than being negatively related as we observed under small spillovers. How does this relate to the d'Aspremont and Jacquemin result in the case of large spillovers? Intuitively, the introduction of large spillovers tends to counter the negative monopolistic effects obtained in a fully cooperative model such as reduced activity levels relative to competitive models. In the case of the noncooperative model, as the level of spillover rises, firms tend to "free-ride" on the other firm's knowledge.[12]

[12] The reader can observe this "free-riding" characteristic by noting that in the noncooperative model firms see their levels of R&D falling, while the fully cooperative firms see their R&D levels rise as the spillover parameter increases.

REFERENCES

d'Aspremont, Claude and Jacquemin, Alexis, "Cooperative and Noncooperative R&D in Duopoly with Spillovers," *American Economic Review*, December 1988, *78*, 1133–37.

Hartwick, John M., "Instability and Specialization in Two-Stage R&D Duopoly Games," Queen's University, mimeo., 1988.

Seade, Jesus K., "The Stability of Cournot Revisited," *Journal of Economic Theory*, August 1980, *23*, 15–27.

Spencer, Barbara J. and Brander, James, "International R&D Rivalry and Industrial Strategy," *Review of Economic Studies*, October 1983, *50*, 707–22.

Cooperative and Noncooperative R&D in Duopoly with Spillovers: Erratum

By CLAUDE D'ASPREMONT AND ALEXIS JACQUEMIN*

In our article published in this *Review* (volume 78, no. 5, December 1988, pp. 1133–37), we have shown that in the presence of sufficient spillovers of the R&D benefits, duopolists, cooperating in R&D but not in output, spend more on R&D than noncooperating firms at both stages, and also produce more output, closest to the socially optimal level. A second symmetric result is that for small spillovers, duopolists cooperating neither in R&D nor in output spend more on R&D and produce more output than cooperative firms. However, this result has been obscured by an obvious inequality inversion and other typos implying a modification in the conclusions. Indeed in fn. 13 on p. 1135 one should have that $\tilde{x} > x^*$ iff

$$\frac{(1+\beta)}{4b\gamma-(1+\beta)^2} > \frac{(2-\beta)}{4.5b\gamma-(2-\beta)(1+\beta)}$$

$$\text{or } \beta > 0.41.$$

This implies that the classification in fn. 16 on p. 1137 should be changed to

$$x^{**} > x^* \geq \tilde{x} > \hat{x}; \qquad Q^{**} > Q^* > \hat{Q} > \tilde{Q}.$$

Therefore, "for some spillovers, such that $\beta \leq 0.4$, the classifications are different *and* the 'second-best' for R&D is obtained by a *non*cooperative behavior in both stages" (p. 1137, lines 5–8).

These results have now been generalized to a wide class of oligopoly models (K. Suzumura, 1989); the stability conditions of the solutions have been established (I. Henriques, this issue); and the effects of various rates of research spillovers have been explored (N. Vonortas, 1989). All these papers emphasize the crucial role played by the spillover parameter β and, implicitly or explicity, suggest the usefulness of endogenizing it. Three main types of factors corresponding to the three dimensions of the relevant game can influence the value of β.

First, there is the nature of the research: a priori, results of precompetitive, generic research are less easily appropriable and therefore lead to more spillovers than those of specific, applied development activities.

Second, the nature of the product is important with the usual distinction between homogeneous and differentiated goods that could lead to different solution concepts such as Cournot and Bertrand: a priori, spillovers are superior in the former case, given that the more standardized is a product, the easier it is to embody in it the results of R&D.

Third, the nature of the contract and the degree of perfection in information also affect the rate of spillovers. On one side, these spillovers are higher for members of the cooperative agreement than for noncooperative firms. On the other side, within the cooperative group itself, β can vary according to the organizational arrangement, $\beta = 1$ corresponding to perfect communication and utilization of the resulting information, for example, through integrated laboratories.

In terms of empirical verification and public-policy decisions, it is therefore crucial to extend the analysis of the welfare effects of cooperative R&D by taking into account the main determinants that can modify, at one moment of time and over time, the level of the corresponding externalities.

REFERENCES

Henriques, I., "Cooperative and Noncooperative R&D in Duopoly with Spillovers:

*Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium.

Comment," *American Economic Review*, June 1990, 80, 638–40.

Suzumura, K., "Cooperative and Noncooperative R&D with Spillovers in Oligopoly," Working Paper, The Institute of Economic Research, Hitotsubashi University, Naka 2-1, Kunitachi, Tokyo, Japan, 1989.

Vonortas, N., "Inter-Firm Cooperation in Imperfectly Appropriable Research: Industry Performance and Welfare Implications," Working Paper, Department of Economics, New York University, 1989.

Deloitte & Touche

Suite 2400
Third National Financial Center
424 Church Street
Nashville, TN 37219-2396

Independent Auditor's Report

Executive Committee
The American Economic Association

We have audited the accompanying balance sheets of The American Economic Association as of December 31, 1989 and 1988, and the related statements of revenues and expenses, changes in general fund and restricted fund balances and cash flows for the years then ended. These financial statements are the responsibility of the Association's management. Our responsibility is to express an opinion on these financial statements based on our audits.

We conducted our audits in accordance with generally accepted auditing standards. Those standards require that we plan and perform the audit to obtain reasonable assurance about whether the financial statements are free of material misstatement. An audit includes examining, on a test basis, evidence supporting the amounts and disclosures in the financial statements. An audit also includes assessing the accounting principles used and significant estimates made by management, as well as evaluating the overall financial statement presentation. We believe that our audits provide a reasonable basis for our opinion.

In our opinion, the financial statements referred to above present fairly, in all material respects, the financial position of The American Economic Association as of December 31, 1989 and 1988, and the results of its operations and its cash flows for the years then ended, in conformity with generally accepted accounting principles.

Deloitte & Touche

Certified Public Accountants
March 2, 1990

THE AMERICAN ECONOMIC ASSOCIATION BALANCE SHEETS FOR THE YEARS ENDED DECEMBER 31, 1989 AND 1988

	1989	1988
Assets		
CASH	$ 649,991	$ 972,272
INVESTMENTS, at market (Notes A and B)	5,756,069	4,968,804
ACCOUNTS RECEIVABLE, no allowance for doubtful accounts considered necessary	52,189	154,146
INVENTORY OF *Index of Economic Articles*, at cost	205,395	178,659
PREPAID EXPENSES	31,448	22,673
OFFICE FURNITURE AND EQUIPMENT—at cost, less accumulated depreciation of $85,151 (1989) and and $77,721 (1988)	59,774	62,236
	$6,754,866	**$6,358,790**
Liabilities and Fund Balances		
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	$ 600,843	$ 538,268
DEFERRED REVENUE (Note A)		
Life membership dues	33,890	36,570
Other membership dues	715,403	649,297
Subscriptions	517,102	548,611
Job Openings for Economists	24,241	23,727
	1,290,636	1,258,205
ACCRUAL FOR SURVEY OF MEMBERS (Note A)	159,051	277,918
FUND BALANCES:		
General	4,641,042	3,966,108
Net worth	4,641,042	3,966,108
Restricted (Note A)	63,294	318,291
Total Fund Balances	4,704,336	4,284,399
	$6,754,866	**$6,358,790**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF REVENUES AND EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1989 AND 1988

	1989	1988
REVENUES FROM DUES AND ACTIVITIES:		
Membership dues and subscriptions	$1,048,752	$ 943,758
Nonmember subscriptions	759,282	756,611
Job Openings for Economists subscriptions	37,498	34,111
Advertising	134,428	128,493
Sale of *Index of Economic Articles*	90,695	175,596
Sale of copies, republications, and handbooks	39,780	44,136
Sale of mailing list	56,591	49,268
Annual meeting	83,426	70,983
Sundry (Exhibit I)	83,296	92,925
Operating Revenues	**2,333,748**	**2,295,881**
PUBLICATION EXPENSES:		
American Economic Review	743,755	708,580
Journal of Economic Literature	950,835	911,429
Survey of Members (Note A)	70,000	70,000
Job Openings for Economists	62,440	58,666
Index of Economic Articles	50,272	96,399
Journal of Economic Perspectives	393,262	361,079
	2,270,564	2,206,153
OPERATING AND ADMINISTRATIVE EXPENSES:		
General and administrative:		
Salaries	205,473	176,021
Rent	22,001	20,388
Other (Exhibit II)	214,814	228,969
Committee	63,894	54,810
Annual meeting	6,608	7,544
	512,790	487,732
Operating Expenses	**2,783,354**	**2,693,885**
Operating Deficit	(449,606)	(398,004)
INVESTMENT INCOME RECOGNIZED (Note B)	297,054	280,512
EXPENSES IN EXCESS OF REVENUES	**($ 152,552)**	**($ 117,492)**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN GENERAL FUND BALANCE

	Total	Operations	Market Value Adjustments
Balance at January 1, 1988	**$3,898,027**	**$2,305,091**	**$1,592,936**
Add change in market value of investments	185,573	–	185,573
Deduct expenses in excess of revenues	(117,492)	(117,492)	–
Balance at December 31, 1988	**3,966,108**	**2,187,599**	**1,778,509**
Add change in market value of investments	827,486	–	827,486
Deduct expenses in excess of revenues	(152,552)	(152,552)	–
Balance at December 31, 1989	**$4,641,042**	**$2,035,047**	**$2,605,995**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN RESTRICTED FUND BALANCE

	Balance at January 1	Receipts	Disbursements	Balance at December 31
YEAR ENDED DECEMBER 31, 1988:				
The Alfred P. Sloan Foundation, Ford Foundation, and Federal Reserve System grants for increase of educational opportunities for minority students in economics	$107,002	$288,965	$207,129	$188,838
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	–	–	5,000
The Andrew W. Mellon Foundation, and Alfred P. Sloan Foundation grants to study economic graduate education in the United States	–	200,000	84,040	115,960
Sundry	2,893	5,600	–	8,493
	$114,895	**$494,565**	**$291,169**	**$318,291**
YEAR ENDED DECEMBER 31, 1989:				
The Alfred P. Sloan Foundation, Ford Foundation, and Federal Reserve System and Rockefeller Foundation grants for the increase of educational opportunities for minority students in economics	$188,838	$115,136	$235,786	$ 68,188
The Andrew W. Mellon Foundation, Alfred P. Sloan Foundation and National Science Foundation grants to study economic graduate education in the United States	115,960	178,886	317,322	(22,476)
The Olin Foundation grant to confer on intellectual property rights	–	12,500	3,011	9,489
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	–	–	5,000
Sundry	8,493	1,100	6,500	3,093
	$318,291	**$307,622**	**$562,619**	**$ 63,294**

The American Economic Association Statements of Cash Flows for the Years Ended December 31, 1989 and 1988

	1989	1988
Cash Flows from Operating Activities:		
Receipt of membership dues and subscriptions	$2,468,136	$2,230,521
Disbursements to suppliers and employees	(2,853,839)	(2,557,537)
Receipts of restricted funds	307,622	494,565
Disbursements made from restricted funds	(562,619)	(291,169)
Net Cash Used in Operating Activities	(640,700)	(123,620)
Cash Flows from Investing Activities:		
Purchase of office furniture and equipment	(18,856)	(11,584)
Purchases of investments	(3,500,195)	(6,459,995)
Proceeds from sale of investments	2,712,930	6,251,160
Proceeds from interest and dividends on investments	1,124,540	466,085
Net Cash Provided by Investing Activities	318,419	245,666
Net (Decrease) Increase in Cash	(322,281)	122,046
Cash at Beginning of Year	972,272	850,226
Cash at End of Year	**$ 649,991**	**$ 972,272**
Reconciliation of Expenses in Excess of Revenues to Net Cash Used in Operating Activities:		
Expenses in excess of revenues	($ 152,552)	($ 117,492)
Adjustments to reconcile expenses in excess of revenues to net cash used in operating activities:		
Depreciation	21,318	15,868
Changes in assets, liabilities, and fund balances:		
Decrease (increase) in accounts receivable	101,957	(123,071)
(Increase) decrease in inventory	(26,736)	27,299
Increase in prepaid expenses	(8,775)	(3,843)
Increase in accounts payable and accrued liabilities	62,575	27,419
Increase in deferred revenue	32,431	57,711
(Decrease) increase in accrual for *Survey of Members*	(118,867)	69,605
Investment income recognized	(297,054)	(280,512)
(Decrease) increase in restricted fund balance	(254,997)	203,396
Total adjustments	(488,148)	(6,128)
Net Cash Used in Operating Activities	**($ 640,700)**	**($ 123,620)**

See notes to financial statements.

The American Economic Association Notes to Financial Statements for the Years Ended December 31, 1989 and 1988

A. Summary of Significant Accounting Policies

Investments are accounted for on a market value basis. The investment income recognized is modified to reflect only the Association's approximate historical average rate of return, which is currently 5%. Investment income represents 5% of the total cash and market value of investments at the beginning of the year. The change in market value of investments and dividends and interest earned net of investment income recognized is recorded directly to the general fund.

Accrual for Survey of Members. Every three to five years the Association publishes a survey which lists, among other things, the names and addresses of its membership. This survey was most recently published in 1989 and distributed at no cost to the membership. In order to properly match the publishing cost of this survey with revenue from membership dues, the Association provided $70,000 in 1989 and in 1988 for estimated publishing costs which will reduce actual survey expenses in the year of publication.

Deferred revenue represents income from membership dues and subscriptions to the various periodicals of the Association which are deferred when received. These amounts are then recognized as income following the distribution of the specified publications to the members and subscribers of the Association. Income from life membership dues is recognized over the estimated average life of these members.

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under Section 501(c) (3) of the Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists. The Association has been determined to be an organization which is not a private foundation.

Certain restricted funds are administered on a reimbursement basis; therefore, disbursements are allowed prior to receipt of grant proceeds.

Certain reclassifications have been made to the 1988 amounts in order to conform to the 1989 presentation.

B. Investments and Investment Income

Investments consist of:

	December 31, 1989		December 31, 1988	
	Cost	Market	Cost	Market
Government obligations, bonds, and commercial paper	$ 586,999	$ 617,179	$ 922,760	$ 954,291
Mutual funds	4,578,432	5,138,890	4,004,113	4,014,513
	$5,165,431	$5,756,069	$4,926,873	$4,968,804

Investment income recognized consists of:

	Year Ended December 31,	
	1989	1988
Government obligations, bonds, and commercial paper—interest	$ 92,817	$ 90,847
Corporate stocks and mutual funds—cash dividends	471,922	218,198
Corporate stocks and mutual funds—net gain on sale	11,094	15,792
Change in market value	548,707	141,248
Transfer to general fund, net	(827,486)	(185,573)
Investment income recognized, net	$297,054	$280,512

C. Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was approximately $42,000 for both 1989 and 1988.

D. Ratio of Net Worth to Expenses

The ratio of net worth at December 31, 1989 to 1990 budgeted expenses is 1.49 and the ratio of net worth at December 31, 1988 to actual 1989 expenses is 1.42.

EXHIBIT I—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF SUNDRY REVENUES FOR THE YEARS ENDED DECEMBER 31, 1989 AND 1988

	1989	1988
AER submission fees	$49,855	$45,860
Royalties	27,833	29,062
CSWEP membership dues	4,721	16,280
Donations	368	529
Permission to reprint	327	200
Foreign postage	192	290
Miscellaneous income	–	704
	$83,296	**$92,925**

EXHIBIT II—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1989 AND 1988

	1989	1988
Dues and subscriptions	$ 59,183	$ 66,405
Mailing list file maintenance	33,446	37,532
Postage	26,501	18,371
Depreciation (straight-line method)	21,318	15,868
Accounting and legal	20,292	20,222
Investment counsel and custodian fees	16,801	18,445
Office supplies	11,547	7,816
President and president-elect expenses	7,227	6,000
Insurance and miscellaneous	6,722	6,302
Election expenses	5,936	10,131
Telephone	4,479	4,298
Bank charges	713	12,207
Travel	649	5,372
	$214,814	**$228,969**

The American Economic Review

B. Studies

ARTICLES

SEPTEMBER 1990

THE AMERICAN ECONOMIC ASSOCIATION

•Printed at Banta Company, Menasha, Wisconsin.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002-8282), September 1990, Vol. 80, No. 4, is published five times a year (March, May, June, September, December). by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: $42.00, $50.40, or $58.80 depending on income. A membership also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add $16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 2014 Broadway, Suite 305, Nashville, TN 37203.

Founded in 1885

Officers

President
GERARD DEBREU
University of California-Berkeley

President-elect
THOMAS C. SCHELLING
Harvard University

Vice-Presidents
RUDIGER W. DORNBUSCH
Massachusetts Institute of Technology
ALLAN H. MELTZER
Carnegie Mellon University

Secretary-Treasurer
C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review
ORLEY C. ASHENFELTER
Princeton University

Editor of The Journal of Economic Literature
JOHN PENCAVEL
Stanford University

Editor of The Journal of Economic Perspectives
JOSEPH E. STIGLITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee
GEORGE A. AKERLOF
University of California-Berkeley
ISABEL V. SAWHILL
Urban Institute
STANLEY FISCHER
World Bank
LAWRENCE H. SUMMERS
Harvard University
GREGORY C. CHOW
Princeton University
SUSAN ROSE-ACKERMAN
Yale University

EX OFFICIO Member
ROBERT EISNER
Northwestern University

GUY HENDERSON ORCUTT

Distinguished Fellow

1989

Guy H. Orcutt's early research contributed to economists' understanding of the problems of testing the significance of relationships among autocorrelated time-series. The difficulties of using aggregative time-series to test theories about the behavior of the economy, however, led Orcutt to develop a new approach to economic modeling. His approach was highly disaggregative; it started with the basic decision units of the economy, households and firms. He has devoted his career to increasing knowledge of the behavior of these microeconomic units and using that knowledge to improve predictions about the operation of the economy and the impact of policy changes.

Orcutt's path-breaking work on micro analytical models of economic behavior, especially in the household sector, stimulated the development of a whole series of models that have helped policymakers estimate the costs and effects of alternative Social Security, income maintenance, and other policy changes. Recent advances in computer technology are making Orcutt's vision of using micro analytic models to enrich understanding of macroeconomic behavior increasingly feasible.

Orcutt has worked tirelessly to increase the availability of survey data, especially longitudinal data needed for better understanding of the behavior of households and firms. He has contributed both to the methodology of social experimentation and to new ways of learning from the natural experiments that arise when policy is changed.

Orcutt's innovative thinking about economic behavior and enthusiasm for micro analytic research and modeling have inspired students and colleagues in many parts of the world.

Guy H. Orcutt

THE AMERICAN ECONOMIC REVIEW

September 1990

VOLUME 80, NUMBER 4

Articles

Shorter Papers

•Submit manuscripts (4 copies), 50 pages maximum, single-sided, double-spaced, to:

Orley Ashenfelter, Editor, *AER*, 209 Nassau Street, Princeton, NJ 08542-4607.

•Submission fee: $50 for members; $100 for nonmembers. Please pay with a check or money order payable in United States Dollars. Canadian and Foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Style guides will be provided upon request.

Editorial Statement

It is the policy of the *American Economic Review* to publish papers only where the data used in the analysis are clearly and precisely documented, are readily available to any researcher for purposes of replication, and where details of the computations sufficient to permit replication are provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary, or if, for some other reason, the above requirements cannot be met.

The Origins of American Industrial Success, 1879–1940

By GAVIN WRIGHT*

The United States became the world's preeminent manufacturing nation at the turn of the twentieth century. This study considers the bases for this success by examining the factor content of trade in manufactured goods. Surprisingly, the most distinctive characteristic of U.S. manufacturing exports was intensity in nonreproducible natural resources; furthermore, this relative intensity was increasing between 1880 and 1920. The study then asks whether resource abundance reflected geological endowment or greater exploitation of geological potential. It was mainly the latter. (JEL 042)

Recent thinking about American economic performance has been marked by alarm over the country's loss of its "competitive edge." Most of this discussion is not rooted in an understanding of the historical origins of the economic leadership now thought to be in jeopardy. Modern economists tend to assume that the American advantage has been technological and dates from the remote recesses of history, about as far back as anyone really cares to go. In a volume on U.S. competitiveness, Harvey Brooks writes: "Both our firms and our government, *long accustomed to being the technological leaders in almost every field*, have until recently measured their performance against domestic rather than foreign competitors" (Bruce R. Scott and George C. Lodge, 1985, p. 331; emphasis added). For one country to maintain a technologically based advantage over others for long historical periods is anomalous, and surely calls for explanation. Indeed, it is difficult to see how policies can respond appropriately to "what we have lost" without knowledge of what it was that we had and how we got it. It would be an understatement, however, to say that the subject has been understudied. This paper makes a modest beginning by analyzing American trade in manufactured goods between 1879 and 1940. The competitive success of American manufacturing exports in foreign markets is by no means a comprehensive measure of "success." But because the turn of the century marked the emergence of the United States to a position of world economic preeminence, we may hope to learn something about the broader questions by studying the characteristics of the country's trade with the rest of the world during that key era.

The results are surprising. They suggest that the single most robust characteristic of American manufacturing exports was intensity in nonreproducible natural resources. In fact, their relative resource intensity was *increasing* over the half-century prior to the Great Depression. This does not mean that there was no American technological leadership, in the broad sense of that term. Abundant resources were themselves in many ways a reflection of the advanced state of American technology. But the distinctively American industrial innovations were in many respects specific to the pre-World War II U.S. resource environment and national market, both of which were unique among the countries of the world. Since then, relative American resource abundance has greatly diminished, not primarily

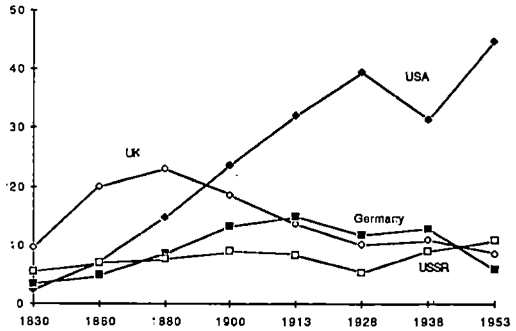
*Department of Economics, Stanford University, Stanford, CA 94305. This research was supported by the Center for Economic Policy Research at Stanford, and the paper was written while in residence at the Institute for Advanced Study in Princeton, NJ. Those who have offered useful advice are too numerous to list, but the author is particularly grateful to Paul David for suggesting and encouraging the project, and to Stanley Engerman, Alex Field, Albert Hirschman, Nate Rosenberg, Gary Saxonhouse, and Robert Staiger for specific suggestions. Research assistance by David Green and Jeff Sundberg was much appreciated.

from depletion of national reserves but because of the integration of world markets for minerals and other commodities. Twentieth-century patterns of resource discovery and production suggest that the historic basis for U.S. mineral abundance was much more a matter of early "development" than of geological "endowment."

I. The Ascendance of American Industry on a Global Scale

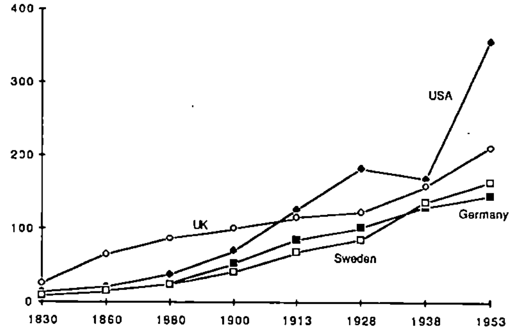
Americans have enjoyed high material living standards since the eighteenth century if not earlier, and the acceleration to modern rates of per capita growth occurred during the first half of the nineteenth century. Broadly based American *industrial* leadership on a worldwide basis, however, can only be dated from the very end of the nineteenth century. According to Paul Bairoch (1982), the U.S. share of total world manufacturing output passed Great Britain's between 1880 and 1900 (Chart 1). In per capita levels of industrial output, the United States was a weak fourth among the nations of the world in 1880, and surpassed Britain only after 1900 (Chart 2). Contemporary testimony suggests that American technology and manufactured goods began to play a qualitatively different role in the world as of the 1890s or shortly thereafter. The first wave of alarmist European books on "Americanization" dates from 1901 and 1902, with titles and themes (*The American Invaders*, 1901; *The Americanization of the World*, 1901; *The American Invasion*, 1902) that would again become familiar in the 1920s and 1960s (William Woodruff, 1975, p. 123). Rapid inflows of standardized, machine-made American shoes after 1894 (said to be more comfortable and more stylish than the traditional types) caused consternation in the British boot-and-shoe industry and forced a drastic technological overhaul (R. A. Church, 1968). Equally dramatic was the burst of American exports of machine tools and other engineering goods after 1895, not only to Britain but to the Continent and other parts of the world (Roderick C. Floud, 1974, pp. 60–62; 1976, pp. 72–82). Though the suddenness of the American "invasion" after 1895 may be attributable to temporary factors, it seems clear that a crossover point of some sort was reached at that time.[1]

Chart 1. Shares of World Industrial Output, 1830–1953



Source: Bairoch (1982, pp. 296, 304).

Chart 2. Industrial Output per Capita



Source: Bairoch (1982, pp. 294, 302).

Industry studies seem to confirm this timing. Robert Allen has shown that prior to the 1890s American blast furnaces had no distinctive world-class status in either labor productivity or fuel efficiency (Allen, 1977, pp. 608–609). By 1900, after key break-

[1] S. J. Nicholas (1980) argues that the apparent decline in the price of American "engineering goods" mainly tracks the prices of iron and steel products, and that the sudden "invasion" of U.S. goods reflected temporary delivery lags by British firms during 1895–1900. As argued below, both of these elements reflected more lasting features of American industrial success.

throughs in adapting the technology to the new Mesabi iron ore, the U.S. industry was the world leader by both of these indicators. Pig iron was an input in the production of steel, which was in turn crucial for railroads, construction, and a wide range of machinery and manufactured goods. According to Allen, before the 1890s American steel rails would not have been competitive in the domestic market without tariff protection (Allen, 1981). Advances in steel were in turn complementary to progress in other industries. U.S. rubber-tire makers, for example, were well behind the French during the bicycle craze of the 1890s, and only gained a productivity advantage in conjunction with mass production of automobiles shortly before World War I (M. J. French, 1987, p. 66). None of this denies that twentieth-century U.S. technology emerged from an evolutionary learning process over a much longer period, as economic historians have long stressed (Paul David, 1975; Nathan Rosenberg, 1976; David Hounshell, 1984). But the qualitative changes in industrial America's place in the world after 1890 justify a closer look at this period.

The timing of U.S. industrial performance corresponds closely to the more comprehensive finding of U.S. world leadership by Angus Maddison, based on estimates of Gross Domestic Product per man-hour (Maddison, 1982, p. 212; compare also Moses Abramovitz, 1986). But Maddison seems to assume that U.S. leadership in *productivity* corresponded closely to a position of "world leadership" in *technology*. This is surely not the only possibility. In terms of conventional growth-accounting, the U.S. edge could equally well have been attributable to capital or natural resources. Interestingly, Maddison's figures actually show that the world leader in GDP per man-hour prior to World War I was not the United States but Australia. His explanation is confined to a footnote: "In defining productivity leadership, I have ignored the special case of Australia, whose impressive achievements before the First World War were due largely to natural resource advantages rather than to technical achievements and the stock of man-made capital" (p. 258). Can we be certain that the United States was not also a "special case" whose performance depended on "natural resource advantages"?

Contrary to expectations of increasing resource scarcity, post-Civil War American development featured *declining* relative costs of materials (Louis P. Cain and Donald G. Paterson, 1981, pp. 358–360). Major new metals discoveries continued until World War I, while the rate of discovery of new oil fields accelerated after 1900 (U.S. National Resources Committee, 1937, p. 149). The timing of leadership in industrial production coincides remarkably with American world leadership in coal production (after 1900), and that margin also grew over time. The United States was also the world's leading producer of copper, petroleum, iron ore, zinc, phosphate, molybdenum, lead, tungsten, and many other minerals. At the same time, continuing advances in internal transportation reduced the real costs to manufacturers, creating what historical geographers call the "minerals-dominant economy" of the late nineteenth century (Harvey S. Perloff and Lowden Wingo Jr., 1961, pp. 193–197). The improvements were often qualitative as well as quantitative, most strikingly perhaps in the iron ore from the rich Mesabi range, which began to arrive in the steel mills of the lower Great Lakes in the 1890s. Allen's estimates of total-factor-productivity in iron and steel as of 1907–1909 put the United States at a par with Germany (15 percent ahead of Britain), but the ratio of horsepower to worker was twice as large in America as in either of the other two contenders (Allen, 1979, p. 919). If we were to adopt the conventional view that resource abundance is an *alternative* to technologically based manufacturing, we might well be led to question the authenticity of America's leadership position before World War II. But as argued below, this is not the only choice available to us.

II. Hypotheses from the International Trade Literature

A number of hypotheses bearing on American industrial history emerge from the literature on the bases fcr international trade. According to the Heckscher-Ohlin model, the composition of a country's trade reflects the relative abundance of factors in that country's endowment. Simple two-factor versions of this theory have frequently been rejected, beginning with the "Leontief Paradox," which revealed that in 1947 U.S. exports were more capital-intensive than were competitive imports (Wassily Leontief, 1953). Attempts to rationalize this result, however, have generated more refined propositions. According to the "neo-factor-proportions" approach, American exports have actually been intensive in skills or human capital. This interpretation was suggested by Leontief himself, and has been supported by an empirical regularity first identified by Irving B. Kravis (1956a), that average wage levels in American export industries have been persistently higher than wage levels in import-competing industries. It has become a standard convention in empirical trade studies to take the relative industry wage as a proxy for skill requirements, and on this basis the skill intensity of American exports has been claimed as a pattern as far back as 1899 if not earlier (Helen Waehrer, 1968). Studies for more recent periods have supported this view with detailed evidence on the occupational structure of the labor force (Donald B. Keesing, 1968).

An alternative "third factor" interpretation for the paradox is that capital is complementary to natural resources, and that the United States had moved into a position of resource scarcity by 1947 (Kravis, 1956b). This possibility is supported by Jaroslav Vanek's important study of the natural resource content of U.S. foreign trade, 1870–1955 (Vanek, 1963), which showed that the country had moved from a net export to a net import position in natural resources over that period. This finding raises the possibility that U.S. comparative advantage may have had a different basis at an earlier time.

A different (though not necessarily mutually exclusive) intellectual strategy is taken by the "neo-technology" approach. The concept of a "technological gap" between the United States and the rest of the world was a commonplace in discussions of trade and direct investment during the 1950s and 1960s (Atlantic Institute, 1970). Though theory makes a sharp distinction between "factor proportions" and "technology" effects, in practice the two ideas are often similar. Employment of skilled professional and scientific personnel is correlated with investment in research, often called "R&D intensity" or simply the "technology factor" (Raymond Vernon, 1970). Similarly, American "technology" has often been linked as much with managerial performance as with science-based production methods. Since the vertically integrated modern business corporation developed earlier and diffused more widely in the United States than elsewhere (Alfred D. Chandler and Herman Daems, 1980), the conceptual correlations among technology, organization, and personnel are likely to be high.

A more difficult conceptual challenge is technological leadership manifest in the form of new products, exported from the United States because they were unavailable elsewhere (Kravis, 1956b). Because exports were small as a percentage of output for almost all American industries, the U.S. case would seem to be a likely example of the historical process described by Staffan Burenstam Linder (1961) whereby new products originally designed for the domestic market begin to enter foreign trade as production expands: "International trade is really nothing but an extension across national frontiers of a country's own web of economic activity" (p. 88). Vernon's "product-cycle" model is perhaps the best-known version: *New products* tended to appear first in the United States because they were responsive to *high-income wants*, and because they were associated with an environment of *high labor costs*. As processes became more mature and routine, trade would be displaced by production abroad, but the volume of U.S. exports was maintained by a continuing flow of new innovations (Vernon, 1966).

There is an ever-present danger of anachronism in applying such concepts historically. The United States did not invent the firearm, the shoe, the bicycle, the camera, or the automobile, and the American versions of these goods were not regarded in European countries as well suited to "high-income wants" (which were better served by the English or French). The size and character of the U.S. domestic market were certainly crucial, but the bulk of the new American exports were producers goods, whose "novelty" lay not so much in consumer taste as in technical specifications or quality. The approach taken here therefore concentrates on the supply side, by analysis of the changing factor content of manufacturing trade over the era of American ascendancy. Though we cannot claim to measure or establish the nature of American "technological leadership" in a rigorous sense, we can illuminate that subject by finding the characteristics of those U.S. products that had the greatest impact on world markets.

This has been the approach of earlier historical work.[2] Using the standard methodology of empirical trade studies, N. F. R. Crafts and Mark Thomas present an analysis of comparative advantage in British manufacturing trade between 1910 and 1935, which they contrast unfavorably with that of the United States (Crafts and Thomas, 1986). They find that Britain continued to export products intensive in capital and unskilled labor and to import goods intensive in human capital (as reflected in the average industry wage). A similar regression for the United States in 1909 shows a reverse result. They conclude: "The U.S. appears already to be following the 'advanced country' pattern of exporting human capital intensive goods and importing unskilled labor-intensive goods in 1909" (p. 637). The next section considers whether this impression should be modified on the basis of a richer data set.

III. New Evidence on American Trade in Manufactures

A. *Average Factor Intensities*

One of the reasons that American manufacturing trade has been understudied is that the Commerce Department trade data are entirely separate from the censuses of manufactures, which have no information about foreign markets. It is not a simple task to match these two sources. Fortunately, a Stanford dissertation by Mary Locke Eysenbach estimated production coefficients for 165 industries according to the system used in Leontief's 1947 interindustry study, and matched these to export and import data for 1879, 1899, and 1914 (Eysenbach, 1976). The present research has replicated her procedures and extended the data set to 1909, 1928, and 1940.[3] For most sample years there are just over 100 usable observations, providing a level of detail roughly comparable to three-digit SITC categories.

To explore the factor intensity of manufacturing trade, I have used Eysenbach's production coefficients to trace relative changes over the entire period of observation. Her capital and labor coefficients are primarily from the census of 1899, while the natural resource coefficients were taken from Vanek (1963) and hence originate in the input-input table for 1947. Thus, this is primarily a study of compositional changes in manufacturing trade over time rather than the actual implicit factor flows in each year. As a sensitivity check, however, estimated coefficients for alternative years have been

[2] An extensive literature on the so-called "labor-scarcity paradox" takes a similar tack, assessing U.S. performance indirectly by measuring the factor-saving bias of U.S. technology relative to British. The suggestion by H. J. Habakkuk (1962) that American technology was capital-intensive and labor-saving has given way to a more complex picture: American methods were more intensive in the use of raw materials and fuel and were characterized by a faster pace and more intensive utilization of capital (David, 1975; Field, 1983). The provocative early successes of the "American system" were limited to a small subset of industries in the 1850s (John James and Jonathan Skinner, 1985). This work concentrates on the mid-nineteenth century, giving little attention to change over time or to the overall scope of U.S. industrial performance.

[3] David Green deserves most of the credit for the detective work that this task entailed.

TABLE 1—CAPITAL-LABOR RATIOS FOR MANUFACTURED GOODS, 1879–1940
($000 PER EMPLOYEE IN 1909 DOLLARS)

	A. 1899 Coefficients					
	1879	1899	1909	1914	1928	1940
Exports	4.186	4.059	4.052	3.961	3.946	3.374
Imports	2.608	2.886	2.785	2.850	2.907	3.221
Exports/Imports	1.61	1.41	1.46	1.39	1.36	1.05
	B. 1909 Coefficients					
	1879	1899	1909	1914	1928	1940
Exports	5.405	4.877	4.967	4.811	4.959	4.193
Imports	2.999	3.079	3.020	3.073	3.486	4.444
Exports/Imports	1.80	1.58	1.64	1.57	1.42	0.94
	C. 1947 Coefficients					
	1879	1899	1909	1914	1928	1940
Exports	4.725	5.170	6.350	6.790	6.330	5.265
Imports	2.910	3.440	3.420	3.690	4.325	5.850
Exports/Imports	1.62	1.50	1.86	1.84	1.46	0.90

Sources: 1899 coefficients from Mary Locke Eysenbach, *American Manufactured Exports*, 1897–1914, New York: Arno Press, 1976, pp. 302–306; 1909 coefficients from U.S. Census of Manufactures; 1947 coefficients form Wassily Leontief, "Factor Proportions and the Structure of American Trade," *Review of Economics and Statistics*, November 1956, *38*, 403–407.

Trade Figures: for 1879, 1899, 1914 from Eysenbach, pp. 271–275; 1909, 1928, 1940 from U.S. Commerce Department, *Foreign Commerce and Navigation of the United States*. Exact industry groupings available on request.

TABLE 2—MEASURES OF SKILL INTENSITY OF MANUFACTURED GOODS, 1879–1940

	A. Percentage Earning More than $12/Week in 1890					
	1879	1899	1909	1914	1928	1940
Exports	52.3	48.7	48.2	45.9	46.6	42.9
Imports	48.5	45.7	47.1	44.1	42.3	41.3
Exports/Imports	1.08	1.07	1.02	1.04	1.10	1.04
	B. Average Wage (1909)					
	1879	1899	1909	1914	1928	1940
Exports	0.467	0.482	0.487	0.502	0.504	0.541
Imports	0.431	0.433	0.460	0.426	0.463	0.471
Exports/Imports	1.09	1.11	1.06	1.18	1.09	1.15
	C. Percentage Women and Child Labor (1909)					
	1879	1899	1909	1914	1928	1940
Exports	10.1	10.7	9.9	11.0	11.2	10.4
Imports	30.6	29.0	30.2	27.8	24.2	21.1
Exports/Imports	0.33	0.37	0.33	0.40	0.46	0.49

Sources: Percent $/week from Eysenbach, pp. 307–311; average wage from 1909 Census of Manufactures (wage bill divided by labor force); women and child labor from 1909 Census of Manufactures (females aged 16 and over, under 16, and males under 16, divided by labor force).

used wherever possible. Since all of the coefficients are U.S.-based, the question of whether the factor content of imports accurately corresponds to foreign production techniques is not addressed. Despite these limitations, the procedures follow the spirit of much of the literature on these subjects, and the results (shown in Tables 1 through 3) are suggestive.

Table 1 does confirm that American manufacturing exports were more capital-intensive than American imports from 1879 to 1928. But in terms of contemporary coefficients, the country's surge to world indus-

TABLE 3—NONRENEWABLE NATURAL RESOURCE COEFFICIENTS IN MANUFACTURING GOODS, 1879-1940 (1947 COEFFICIENTS)

	1879	1899	1909	1914	1928	1940
A. Direct Use						
Exports	0.0742	0.0677	0.0918	0.0988	0.09984	0.0564
Imports	0.0131	0.0194	0.0170	0.0133	0.0290	0.0369
Exports/Imports	5.66	3.49	5.40	7.43	3.39	1.53
B. Direct and Indirect Use						
Exports	0.1107	0.1239	0.1647	0.1800	0.1635	0.1240
Imports	0.0565	0.0747	0.0766	0.0749	0.0934	0.1127
Exports/Imports	1.96	1.66	2.15	2.40	1.75	1.10

Sources: Coefficients from Eysenbach, pp. 297-301; trade figures, see Table 1.

trial supremacy was not marked by a shift toward capital-intensive manufacturing exports, nor by an increasing tendency to trade capital-intensive for labor-intensive manufactures with the rest of the world. (It is interesting that the relative capital intensity of exports *in terms of 1947 coefficients* did rise until 1914, after which it declined.) Movement in the direction of the Leontief Paradox within manufacturing is detectable, at least after World War I.

It should be noted that the figures in Table 1 omit refined sugar, an industry that if included would single-handedly generate a Leontief Paradox for manufacturing in every sample year. If classified as a manufactured good (following Eysenbach), refined sugar would account for nearly one-quarter of manufacturing imports before 1900, and sugar refining (in the United States, at any rate) had a capital-labor ratio five times as high as the average for manufacturing. It is open to question whether sugar refining techniques outside the United States were really this capital-intensive. Because the industry is exceptional and because we are not in any case trying to account for all international flows, it seems more informative to leave it out. Though extreme, sugar refining does illustrate one of the compositional reasons for the trend shown in the first two panels of Table 1, namely, the high capital intensity of many agricultural processing industries, which were declining in relative prominence among U.S. exports. Two of the largest contributors to the decline in relative capital intensity of exports were grain mill products, and meat packing and wholesale poultry.

Table 2 displays two indices of skill intensity: (1) following Eysenbach, the percentage of the labor force earning more than $12 per week in 1890, and (2) the average industry wage in 1909.[4] By both measures, there is some tendency for export industries to pay higher wages than import-competeing industries. But there is little sign of a trend in the relative skill intensity of exports and imports. As measured by the 1890 "high-wage" index, the skill content of exports went steadily downward. As measured by the 1909 average wage, however, the skill content of exports had an upward trend. There was also an upward trend, however, in the skill content of imports by the same measure (excepting 1914). One of the reasons for this puzzling pattern is suggested by the third panel of Table 2, which reveals a much more dramatic contrast between exports and imports in the percentage of the labor force who are women and children (under the age of 16). It is perhaps not surprising to see that imports are far more women-and-child-intensive than exports, since these workers are associated with "low-wage" and labor-intensive processes (but it is interesting that this direct measure of labor-force composition is a clearer sepa-

[4] Several other skill indices were proposed by Eysenbach, all based on 1890 data. They give results similar to those presented here.

rator than capital-intensity or wage levels, which one might take to be more fundamental). What is striking is the decline over time in this relative intensity, entirely concentrated on the import side. Here we have another likely contributor to the trend toward the Leontief Paradox. Employment of women and child workers in American manufacturing was concentrated in only a handful of industries: canning, preserving, and freezing on the one hand, and textiles and apparel on the other. The first remained a strong net export category, but in the second, the growth of imports was increasingly stifled by tariff barriers, particularly after the 1922 Fordney-McCumber tariff.

Easily the largest factor-intensity differentials were in nonreproducible natural resources, as shown in Table 3. Recall that these are weighted averages for manufactured goods alone and exclude entirely exports of agricultural goods and crude materials. We still find not only that U.S. exports had far higher natural resource content than imports but that this trend was growing both absolutely and relatively over *precisely the historical period when the country was moving into a position of world industrial preeminence.* Using the more inclusive index of direct and indirect use, the resource intensity of manufacturing exports grew by 64 percent to its peak, and even after a slight decline, the 1928 level was still nearly 50 percent higher than that of 1879. The figures confirm a little-noticed analysis by Robert E. Lipsey (1963): "The composition of manufacturing exports has been changing ceaselessly since 1879 in a fairly consistent direction—*away from products of animal or vegetable origin and toward those of mineral origin*" (p. 59; emphasis added).

Table 3 also clearly shows that the resource intensity of imports was growing as well, and that signs of a reversal in the relative balance are detectable even in 1928. By 1940, the historic U.S. specialization had virtually disappeared. This is the modern trend identified first and most clearly by Vanek (1963), of no small importance for interpreting recent American industrial history. But because of his choice of dates and coverage, Vanek missed the fact that the declining phase had been preceded by a long epoch of rising natural resource intensity, of no less importance in interpreting the country's place in the industrial world.

B. *Regression Analysis*

Simple factor-intensity comparison between exports and imports is not conclusive in the presence of more than two factors (Edward Leamer, 1980). An apparent pattern of specialization may merely represent the effect of a third factor, acting as a complement or substitute for one of the other two. This section therefore follows the general format of Crafts and Thomas (1986) and earlier studies in the international trade literature by regressing the net trade balance for each industry against measures of factor intensity. On no account should the coefficients be viewed as structural estimates within a Heckscher-Ohlin framework (compare Leamer and Harry P. Bowen, 1981). They are best considered as descriptive summaries of trade patterns in a multifactor setting, a way of pointing out areas of distinctive strength and tracking changes over time. Because the industry or commodity groupings are inevitably arbitrary, R^2 levels by themselves are not particularly meaningful; but t-tests on individual coefficients are a reasonable standard for confidence in that factor's contribution, and R^2 comparisons across years should reflect changes in the tightness-of-fit according to factor content. Following Crafts and Thomas, all reported standard errors were recomputed according to the procedure suggested by Hal White (1980) to adjust for heteroskedasticity in the error structure. The effect generally is to reduce the larger t-ratios, so that what is reported here is a conservative version of the account that leaps from the data using ordinary-least-squares. The results are robust to changes in precise variable definitions and to transformations of the coefficients into factor shares at various discount rates. Trade values have been deflated by export and import price indices (Lipsey, 1963, pp. 142–143; 1913 = 100) so that coefficients may be compared across years.

TABLE 4—REGRESSIONS FOR MANUFACTURED NET EXPORTS OF THE UNITED STATES, 1879–1940

	Constant	Capital/ Labor	Natural Resource Coefficient	Average Wage	Percent Women and Children	R^2
1879	−3127	2092**	−10830	−1853		0.079
	(0.68)	(2.24)	(0.74)	(0.27)		
	−228	1725*	−12690		−156	0.103
	(0.06)	(1.77)	(0.83)		(1.53)	
1899	−4068	3729*	−4324	−802		0.075
	(0.66)	(1.73)	(0.11)	(0.07)		
	1735	3140	−8727		−255**	0.093
	(0.28)	(1.46)	(0.21)		(2.02)	
1909	−8965	2648	46950	959		0.146
	(0.92)	(1.17)	(1.17)	(0.06)		
	260	1810	44154		−380**	0.193
	(0.04)	(0.75)	(0.99)		(2.25)	
1914	−21041**	1600	103103*	28468**		0.261
	(2.56)	(0.53)	(1.71)	(2.12)		
	216	1038	98271*		−329*	0.275
	(0.02)	(0.33)	(1.55)		(1.93)	
1928	−21067	5040	112264**	18856		0.143
	(1.20)	(0.83)	(2.19)	(0.52)		
	−4342	4413	107406**		−333	0.149
	(0.17)	(0.67)	(2.01)		(0.87)	
1940	−31898	−1862	126449**	85642		0.085
	(1.13)	(0.42)	(2.22)	(1.38)		
	23714	−2750	117138**		−629*	0.077
	(1.24)	(0.58)	(2.11)		(1.79)	

Notes: Method of estimation is ordinary least-squares, *t*-ratios (in parentheses) adjusted for heteroscedasticity following procedure of White (1980). *Denotes statistical significance at the 5 percent confidence level; **denotes the 1 percent confidence level. There are 64 nonzero observations in 1879, 83 in 1899, and 96 in the remaining years.

The results in Table 4 are broadly consistent with those of the previous section. The capital-labor coefficient is significant in 1879, but it becomes steadily less so in subsequent years and is actually negative by 1940. Thus indications that the Leontief Paradox emerged historically are still present in a multivariate setting. The natural resource coefficient, on the other hand, begins negative and becomes significantly positive after 1909, reaching its peak (in both level and significance) in 1928.

The coefficients of the two labor force variables are also interesting. The coefficient of the average wage is significantly positive in only one year (1914). The coefficient on the percentage of women and child laborers, by contrast, is significantly negative in four of the six years and nearly so in the remaining two. When both variables are included (not shown), the coefficient on the average wage is negative or insignificant in every year. Furthermore, there is an evident inverse relationship between natural resource intensity and the presence of women and children. It appears, therefore, that the concentration of American net exports in "high wage" industries early in the century was attributable to the absence of women and child workers in these "heavy" industries.[5]

An important amendment to this account emerges from Table 5, which uses a new

[5] This does not necessarily mean that the effect is purely compositional, that is, directly explained by the lower wages paid to women and children. Men who worked in these occupational-industrial categories also received lower wages. But these wages did not reflect "skill" levels so much as the ease with which women and children could be substituted for men in these industries.

TABLE 5—REGRESSIONS FOR MANUFACTURED NET EXPORTS OF THE U.S., 1879–1940

	Constant	Capital and Natural Resources/Labor	Average Wage	Percentage Women and Children	R^2
1879	236	2741**	977		0.058
	(0.05)	(2.17)	(0.14)		
	3815	2234		−182*	0.095
	(1.31)	(1.54)		(1.88)	
1899	2495	5650**	4617		0.057
	(0.32)	(2.81)	(0.40)		
	10015*	4677*		−314**	0.088
	(1.98)	(1.95)		(2.58)	
1909	−2974	9312**	6052		0.165
	(0.31)	(3.46)	(0.37)		
	6955*	8045**		−428*	0.229
	(1.93)	(2.68)		(2.67)	
1914	−15799**	13279**	33918**		0.299
	(2.08)	(3.50)	(2.57)		
	7317**	12198**		−386**	0.321
	(2.23)	(3.07)		(2.68)	
1928	−10667	24084**	28310		0.241
	(0.75)	(2.87)	(0.88)		
	9857	22954**		−399	0.252
	(1.09)	(2.61)		(1.40)	
1940	−33084	12118**	86974		0.095
	(1.14)	(2.23)	(1.36)		
	19478**	10590**		−575	0.083
	(2.00)	(1.89)		(1.87)	

Note: See Table 4.

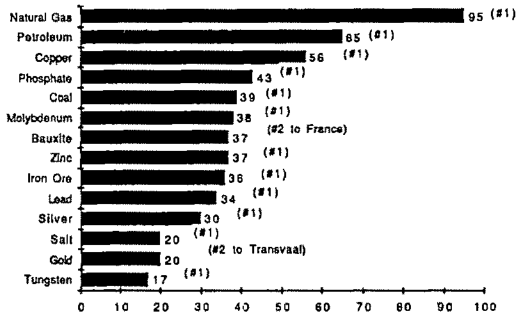
variable created by multiplying the capital-labor ratio and the natural resource coefficient. The results strongly imply that capital and natural resources were complementary factors of production. The coefficient of the new variable is positive through the entire period, growing steadily larger and more significant through 1928. Comparison of R^2 levels between Tables 4 and 5 shows that this new interactive variable is more powerful in accounting for net export performance than the combined effect of its two components, entered separately. The strongest effects are found in 1914 and 1928; in the latter year, for example, the R^2 rises from 0.149 to 0.252 merely by substituting a single variable, the product, for the original two.

This result should caution us against a too-hasty and too-complete rejection of "capital intensity" as a characteristic of American industry. The suggestion is, however, that capital intensity derived not from economy wide abundance of capital per se, but from specialization in an industrial technology in which capital was complementary to natural resources. Strictly speaking, these sorts of tests only describe the direction of trade, not the overall "success" of American industry. But the coincidence of timing between resource intensity and American industrial ascendance obliges us to consider the proposition that the abundance of industrial minerals was a deeper cause of American industry's distinctive strength.

IV. Natural Resources and American Industrial Success

Since industrial success like other historical outcomes requires an uncountable number of mutually interdependent elements, do natural resources really deserve special attention? The scope of America's world leadership in natural resources is displayed in Chart 3, which shows U.S. production of 14 major industrial minerals as a percentage

CHART 3. U.S. MINERAL OUTPUT, 1913: PERCENTAGE OF WORLD TOTAL



Source: Smith (1919), using data from U.S. Geological Survey (1913).

of world totals in 1913. The 95 percent of world natural gas and 65 percent of world petroleum were perhaps of somewhat less economic moment in 1913 than they would be at a later date. But copper, coal, zinc, iron ore, lead, and other minerals were at the core of industrial technology for that era, and in every single case the United States was the world's leading producer by a wide margin. In an era of high transport costs, the country was *uniquely* situated with respect to almost every one of these minerals. Even this understates the matter. Being the number one producer in one or another mineral category is less important than the fact that the *range* of mineral resources abundantly available in the United States was far wider than that in any other country. Surely the link between this geographical status and the world success of American industry is more than incidental. Cain and Paterson (1986) find that between 1850 and 1919, material-using technological biases were significant in nine of twenty American sectors, including those with the strongest export performance, such as petroleum, metals, and machinery.

Resource abundance was a background ingredient in many other distinctively American industrial developments. Continuous-process, mass-production methods, closely associated with modern forms of corporate organization in the analysis of Chandler (1977), were characterized by "high throughput" of fuel and raw materials relative to labor and production facilities (compare Michael Piore and Charles Sabel, 1984). Oliver Williamson (1980) notes that cheap, reliable sources of energy and heat were crucial to this development. Coal was of strategic early importance as a direct source of heat and power, and at a later point as a source of thermal energy for electricity, essential to the efficiency of the moving assembly line and other quasi-flow processes. Alex Field (1987) points out that organizational innovations of this type may be considered "capital-saving" overall, even though firm-level capital requirements were high. In export markets, contemporary comments emphasized non-price competition and particularly the short delivery lags on the part of U.S. suppliers (Nicholas, 1980, pp. 581–587). Quick delivery is a feature one would expect to see where exports have a "vent-for-surplus" quality, because of the length of a production run on a standardized item. In addition, American producer and consumer goods were often specifically designed for a resource-abundant environment. Some of the adjustment problems of U.S. auto companies in recent years stem from their decades of specialization on large, fuel-using cars. There was a parallel problem facing U.S. locomotive manufacturers in the 1920s, who found their foreign sales handicapped by their design for standard-gauge rails, heavy motive power, and heavy train loads (*Markets of the United States*, p. 71).

The emergence of cheap American steel at the end of the nineteenth century was particularly important. Whereas S. J. Nicholas (1980) suggested that the fall in relative U.S. machinery prices was misleadingly proxied by iron and steel prices, it may be that the world success of American engineering goods was buoyed by exactly that development. Table 6 shows the major role played by iron and steel exports over the half-century under discussion. If we aggregate the three headings under which iron and steel products were listed, we find that their share of U.S. manufacturing exports grew steadily, from 5.5 percent in 1879 to 37.5 percent in 1929. If we add in one other

TABLE 6—SHARES OF UNITED STATES MANUFACTURING EXPORTS, 1879–1929 (PERCENT)

	Iron and Steel Products (except Machinery and Vehicles)	Machinery	Automobiles and Parts	SUM (1,2,3)	Petroleum Products	SUM (1,2,3,5)
1879	2.1	3.4	–	5.5	12.1	17.6
1889	2.4	6.1	–	8.5	13.3	21.8
1899	7.6	10.7	–	18.3	9.2	27.5
1913	10.9	14.5	2.3	27.7	10.1	37.8
1923	8.8	12.4	6.4	27.6	13.1	40.7
1926	5.6	12.9	11.5	30.0	16.8	46.8
1927	5.1	13.9	13.3	32.3	14.7	47.0
1928	5.3	16.4	15.7	37.5	13.9	51.4
1929	5.4	16.4	15.7	37.5	13.9	51.4

Source: 1879–1923 (1963), Tables A-8 and A-12; 1926–1929, U.S. Department of Commerce, *Foreign Commerce and Navigation of the United States for the Calendar Year 1929*, Vol. 1, Tables XII and XXIV.

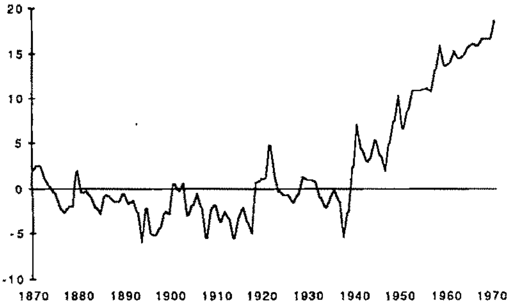
heading in which resource abundance was evidently important, petroleum products, we find that by late 1920s, we have accounted for more than half of all American manufacturing exports. The union of these two sectors is, in essence, the automobile industry. The United States was unquestionably the world's technological leader in automobile production during the 1920s. At the same time, American producers had enormous cost advantages over competitors in raw materials, especially steel. Ford UK faced steel input prices that were higher by 50 percent or more than those paid by the parent company (James Foreman-Peck, 1982, p. 874). It was not accidental that Leontief chose motor vehicles as his most prominently displayed example of the economy as an intricate input-output machine: each million dollars worth of automobiles in 1947 "contained" nearly half that much value in iron and steel, nonferrous metals, and other fabricated metal products (Leontief, 1953, p. 334).

We may also conjecture that there were links between the economy of high throughput and the intensity of the work pace, which also seems to have been a distinctive feature of U.S. industry (Clark, 1987). American firms paid the world's highest real wages and apparently extracted greater effort from the labor force in return. But it is an anachronism to associate "high wages" with "high skill" technologies for the era in which the United States surged to world industrial preeminence. The United States was a well-educated country, but most of the workers in the fast-paced, heavy-industry, mass-production manufacturing in which the country led the world were not well-educated native-born Americans. In 1910 the foreign born and sons of foreign born were more than 60 percent of the machine operatives in the country, and more than two-thirds of the laborers in mining and manufacturing (U.S. Senate, 1911, pp. 332, 334). There is no reason to believe that this labor force was particularly well educated by world standards. Key industries like iron and steel and motor vehicles paid high wages to unskilled workers (who were nonetheless much cheaper than the skilled craft workers used with older technologies), presumably because it was rough, disagreeable, dangerous, demanding work, and because it was vital to have an ample excess labor supply available (compare Daniel Raff, 1988). In the 1930s these industries were central to the movement for industrial unionism, which subsequently provided an alternative mechanism for the continued association between high-wage industries and American industrial success.

V. What Became of American Resource Abundance?

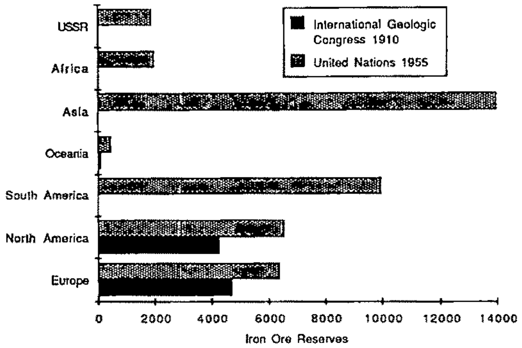
The marked changes in coefficients for 1940 seem to portend the post–World War II pattern, when the United States moved

CHART 4. U. S. NET MINERAL IMPORTS AS A PERCENTAGE OF CONSUMPTION



Source: Manthy (1978, Tables MC1 and MC2).

CHART 5. WORLD IRON ORE RESERVES, 1910 AND 1955



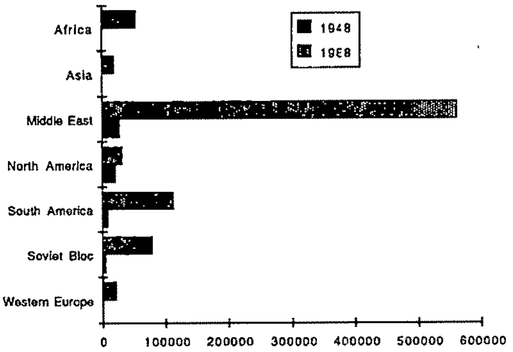
Sources: International Geologic Congress (1910, pp. 1–56): "Actual Reserves" in millions of tons of metallic iron; United Nations (1955, pp. 19–34): "Reserves" in millions of tons of iron content.

steadily and increasingly into a position of net mineral imports (Chart 4). Beginning mainly in the 1920s, one important mineral after another began to enter the net import column: nonferrous metals, bauxite, lead, zinc, copper, iron ore, and petroleum among others. Without conducting extensive global cost comparisons, it is evident that a country for whom resource prices are determined at the margin by imports is not going to have a major locational advantage in resource costs over its industrial rivals. But what exactly was the process of change in America's resource position? A popular conception is that the country has largely exhausted its resource endowment and has had to import so as to avert domestic shortages. Kindleberger has proposed a weaker version of this scenario within the Heckscher-Ohlin framework, in which the more rapid growth of labor and capital relative to resources has turned the country from a net-export to a net-import position with respect to resources (Charles P. Kindleberger 1960, pp. 347–348). It is doubtful that this account is generally valid. Indeed, a closer look at the trend in world mineral supplies casts a different light on the character of the original position.

In 1919 it could confidently be written that "the United States is more richly endowed with mineral wealth than any other country" (George Otis Smith, 1919, p. 282), and such a statement was consistent with the best geological and industrial knowledge of the day. But the clear pattern of discoveries since that time indicates that there was a systematic historical bias in these perceptions, in that American resources had been much more thoroughly explored and exploited than those of other parts of the world. Chart 5 illustrates this process, by comparing world iron ore "reserves", as indicated by a 1910 survey by the International Geologic Congress, with those reported in a United Nations survey in 1955. Granted that quality differences and extraction and transport costs are neglected in a simple chart, nonetheless the pattern is so clear as to be beyond dispute. Europe and North America had by far the largest reserves in 1910, but their "endowments" (which, to be sure, had increased and not decreased) had grown only slightly in the intervening 45 years. What had been a dominating advantage in 1910 was no more than a respectable presence in 1955.

The case of petroleum is even more extreme (Chart 6). Recall that the United States in 1913 (and for a half-century before) had been the world's largest petroleum producer and exporter, by a wide margin. As Chart 6 shows, as late as 1948, North American reserves were nearly equal to those of the Middle East. In 1988, though again reserves of all areas had increased, North America was a minor part of the world petroleum picture. It is difficult to

CHART 6. WORLD CRUDE OIL RESERVES, 1948 AND 1988



Source: American Petroleum Institute (1988, Section II, Table 1): "Estimated Proved Reserves of Crude Oil Annually as of January 1 (millions of barrels)."

avoid the inference that mineral supplies were more a matter of "development" than "endowment."

Where world geological surveys are not available, similar conclusions can be reached by other routes. In the case of bauxite, which takes its name from the French village where it was first developed, the United States and France alternated as first and second in the world until the 1950s. With discoveries in the West Indies in the 1950s, Jamaica quickly moved into first place, at annual production levels larger than those ever achieved in either France or the United States, despite the fact that production levels in those two countries did not decline but continued to grow to levels higher than they themselves had ever achieved. In the late 1960s, Australia replaced Jamaica as number one, again setting new production records without causing an absolute decline in any of the older countries. In both Jamaica and Australia, bauxite production was negligible before World War II. Since the real price of bauxite has declined, it is not the case that domestic reserves have been "exhausted" or that distant supplies have simply been coaxed out along a world supply curve. Rather, early discoveries and mining took place in areas proximate to the early centers of industrial and technological development and within the boundaries of their national jurisdiction.

The last phrase points toward another sense in which resource abundance was historically rather than geologically determined. The United States was the world's largest mineral producing nation, but it was also one of the world's largest countries! Even without Alaska, at 3.5 million square miles, the United States is twice the size of all the countries of eastern and western Europe and Scandinavia combined (excluding Russia). Yet coal and iron ore production in Europe was 30 to 50 percent higher than the U.S. total in the 1910–1913 era. If coal and iron were the imperatives of industrial location ca. 1900, a hypothetical United States of Europe would have rivaled America.

More important than sheer geographic size is economic distance. The United States was a vast free trade area for internal commerce, and the opportunities created by this status provided the incentive for massive investment in transportation infrastructure, including the highly efficient lake transport system that linked Mesabi ore to Pennsylvania coal. In the case of copper, only the combination of national size and efficient internal transportation allow use to say that the "same" economy retained world leadership across the period of American industrial ascendancy, since the early production center in Michigan gave way to remote but larger and richer locations in the Mountain and Southwest regions between 1870 and 1930.

The argument does not stop with national size and efficient transportation. The process of mineral discovery and development was also a prime outlet for creative energies and innovations, often at high levels of technical and organizational sophistication. The United States Geological Survey, formed in 1879 by consolidation of several existing federal surveys, had intimate links with the mining industry. Reports by government geologists in Colorado in the 1880s were crucial in encouraging mining activity and adapting metallurgical knowledge to local requirements (Rodman Wilson Paul, 1960). The American Institute of Mining Engineers became the first speciality group to break away from the American Society of

Civil Engineers. Scientifically trained personnel were also important in expanding the range of *uses* for available minerals. An early report by Yale geologist Benjamin Silliman, Jr., foresaw the commercial possibilities of "cracking" petroleum into various compounds, opening up arrays of new uses for what had been considered a useless waste material (Robert V. Bruce, 1987, pp. 140–142). But as Nathan Rosenberg (1985) points out, much of the early use of science by American industry did not deploy new knowledge at the scientific "frontier," but involved repetitive procedures (such as grading and testing materials) for which scientific training was needed but where the learning was specific to the materials at hand. The abundance of mineral resources, in other words, was itself an outgrowth of America's technological progress.

This view of the matter suggests the answer to the question posed above. The country has not become "resource poor" relative to others, but the unification of world commodity markets (through transport cost reductions and elimination of trade barriers) has largely cut the link between domestic resources and domestic industries. American corporations and engineers have been in the forefront of the globalization of the mineral economy. In essence, the process by which the United States became a unified "economy" in the nineteenth century has been extended to the world as a whole. To a degree, natural resources have become commodities rather than part of the "factor endowment" of individual countries.[6] Presumably this is why international economists now distinguish resource-based "Ricardo goods" from others and treat them separately (for example, Robert Stern and Keith E. Maskus, 1981). This procedure may be appropriate for the contemporary world, but it would be hard to do justice to the historic success of American industry within this conception.

VI. Conclusion

Why has the importance of mineral resources in American industrial history been underappreciated? Concern for the future of natural resources is an ancient theme in economics, but most of the attention has been channeled into two rather different issues: fear of rising costs from increased resource scarcity and fear of national strategic inadequacy in the event of war. Refuting the first fear has long been the economist's favorite pastime, as it has been easy to show that producers substitute away from relatively scarce resources and that the real prices of "nonrenewable" resources have historically declined. The second fear has always seemed noneconomic in character, if not indeed a cooked-up rationalization for subsidy or protection. Having thus dealt with the "problem" of resource exhaustion, it was easy to overlook a logically distinct aspect of the matter: the contribution of resource location to the competitive potential of a country's industries. Some economic historians, to be sure, have long analyzed national economic histories in terms of world geographical patterns (William N. Parker, 1984). But it is perhaps understandable that Americans have not been inclined to attribute their country's industrial success to what appear to be accidental or fortuitous geographic circumstances. Another reason is that American industrial leadership took on a rather different shape after World War II. Over the course of the twentieth century, the country was able to parley its resource-based industrial prosperity into a well-educated labor force, an increasingly sophisticated science-based technology, and world leadership in scientific research itself. In the wake of World War II, there were no serious international rivals in such a wide range of industries that it was easy to lose sight of the resource dimension of industrial performance. After the war, there was a brief period of concern that the nation's resource position had been eroded, culminating in the Paley Commission Report of 1952. But such doubts and fears were largely swept away in the American-led world prosperity of the next 25 years.

[6] Wilfred J. Ethier and Lars E. O. Svensson (1986) show that in a Heckscher-Ohlin framework with mobility of some factors a country's trade pattern in goods is affected only by its endowment of *nontraded* factors.

To be clear about the argument, there is no iron law associating natural resource abundance with national industrial strength. But the distinctive *American* technologies have, as a matter of history, been relatively resource-using. We have now moved from an era in which the rest of the world adapted to an American technology, with varying degrees of difficulty, to an era in which U.S. firms have had to do the adjusting. The adjustment is not made much easier by the consideration developed in this paper, that historical resource abundance was itself largely an outgrowth of American industrial success.

REFERENCES

Abramovitz, Moses, "Catching Up, Forging Ahead, and Falling Behind," *Journal of Economic History*, June 1986, *46*, 385–406.

Allen, Robert, "The Peculiar Productivity History of American Blast Furnaces, 1840–1913," *Journal of Economic History*, September 1977, *37*, 605–33.

______, "International Competition in Iron and Steel, 1850–1913," *Journal of Economic History*, December 1979, *39*, 911–37.

______, "Accounting for Price Changes: American Steel Rails, 1879–1910," *Journal of Political Economy*, June 1981, *89*, 512–28.

Bairoch, Paul, "International Industrialization Levels from 1750 to 1980," *Journal of European Economic History*, Spring 1982, *11*, 269–310.

Bruce, Robert V., *The Launching of Modern American Science 1846–1876*. New York: Knopf, 1987.

Cain, Louis P. and Paterson, Donald G., "Factor Biases and Technical Change in Manufacturing: The American System, 1850–1919," *Journal of Economic History*, June 1981, *41*, 341–60.

______, **and** ______, "Biased Technical Change, Scale and Factor Substitution in American Industry, 1850–1919," *Journal of Economic History*, March 1986, *46*, 153–64.

Chandler, Alfred D., *The Visible Hand*. Cambridge, MA: Belknap Press, 1977.

______ **and Daems, Herman,** eds., *Managerial Hierarchies*. Cambridge, MA: Harvard University Press, 1980.

Church, R. A., "The Effect of the American Import Invasion on the British Boot and Shoe Industry, 1885–1914," *Journal of Economic History*, June 1968, *28*, 223–54.

Clark, Gregory, "Why Isn't the Whole World Developed? Lessons from the Cotton Mills," *Journal of Economic History*, March 1987, *47*, 141–74.

Crafts, N. F. R. and Thomas, Mark, "Comparative Advantage in UK Manufacturing Trade, 1910–1935," *Economic Journal*, September 1986, *96*, 629–45.

David, Paul A., *Technical Choice, Innovation, and Economic Growth*. New York: Cambridge University Press, 1975.

Eckes, Alfred E., *The United States and the Global Struggle for Minerals*. Austin: University of Texas Press, 1979.

Ethier, Wilfred J. and Svensson, Lars E. O., "The Theorems of International Trade with Factor Mobility," *Journal of International Economics*, February 1986, *20*, 21–42.

Eysenbach, Mary Locke, *American Manufactured Exports 1897–1914*. New York: Arno Press, 1976.

Field, Alexander, "Land Abundance, Interest/Profit Rates and Nineteenth-Century American and British Technology," *Journal of Economic History*, June 1983, *42*, 405–31.

______, "Modern Business Enterprise as a Capital-Saving Innovation," *Journal of Economic History*, June 1987, *46*, 473–85.

Floud, Roderick C., "The Adolescence of American Engineering Competition, 1860–1900," *Economic History Review*, February 1974, *28*, 57–71.

______, *The British Machine Tool Industry, 1850–1914*. Cambridge: Cambridge University Press, 1976.

Foreman-Peck, James, "The American Challenge of the Twenties: Multinationals and the European Motor Industry," *Journal of Economic History*, December 1982, *42*, 865–81.

French, M. J., "The Emergence of a U.S. Multinational Enterprise: The Goodyear

Tire and Rubber Company, 1910–1939," *Economic History Review*, February 1987, *40*, 64–79.

Habakkuk, H. J., *American and British Technology in the Nineteenth Century*. Cambridge: Cambridge University Press, 1962.

Hounshell, David, *From the American System to Mass Production, 1800–1932*. Baltimore: Johns Hopkins University Press, 1984.

James, John, and Skinner, Jonathan, "The Resolution of the Labor-Scarcity Paradox," *Journal of Economic History*, September 1985, *45*, 513–40.

Keesing, Donald B., "Labor Skills and the Structure of Trade in Manufactures," in Peter B. Kenen and Roger Lawrence, eds., *The Open Economy*. New York: Columbia University Press, 1968.

Kindleberger, Charles P., "International Trade and the United States Experience, 1870–1955," in Ralph E. Freeman, ed., *Postwar Economic Trends in the United States*. New York: Harper, 1960, pp. 337–73.

Kravis, Irving B., (1956a) "Wages and Foreign Trade," *Review of Economics and Statistics*, February 1956, *38*, 14–30.

______, (1956b) "'Availability' and Other Influences on the Commodity Composition of Trade," *Journal of Political Economy*, April 1956, *64*, 143–55.

Leamer, Edward, "The Leontief Paradox Reconsidered," *Journal of Political Economy*, June 1980, *88*, 495–503.

______ **and Bowen, Harry P.,** "Cross-Section Tests of the Heckscher-Ohlin Theorem: A Comment," *American Economic Review*, December 1981, *71*, 1040–43.

Leontief, Wassily, "Domestic Production and Foreign Trade: The American Capital Position Reexamined," *Proceedings of the American Philosophical Society*, September 1953, *97*, 332–49.

______, "Factor Proportions and the Structure of American Trade," *Review of Economics and Statistics*, February 1956, *38*, 386–407.

Linder, Staffan Burenstam, *An Essay on Trade and Transformation*. New York: Wiley & Sons, 1961.

Lipsey, Robert E., *Price and Quantity Trends in the Foreign Trade of the United States*, Princeton, NJ: Princeton University Press, 1963.

Maddison, Angus, *Phases of Capitalist Development*, Oxford: Oxford Univesity Press, 1982.

Manthy, Robert S., *Natural Resource Commodities: A Century of Statistics*. Baltimore: Johns Hopkins University Press, 1978.

Nicholas, S. J., "The American Export Invasion of Britain: The Case of the Engineering Industry, 1870–1914," *Technology and Culture*, October 1980, *21*, 570–88.

Parker, William N., *Europe, America, and the Wider World*, Vol. 1, Cambridge: Cambridge University Press, 1984.

Paul, Rodman Wilson, "Colorado as a Pioneer of Science in the Mining West," *Mississippi Valley Historical Review*, June 1969, *47*, 34–50.

Perloff, Harvey S. and Wingo, Lowdon, Jr., "Natural Resource Endowment and Regional Economic Growth," in Joseph Spengler, ed., *Natural Resources and Economic Growth*. Washington: Resources for the Future, 1961.

Piore, Michael and Sabel, Charles F., *The Second Industrial Divide*. New York: Basic Books, 1984.

Raff, Daniel M. G., "Wage Determination Theory and the Five-Dollar Day at Ford," *Journal of Economic History*, June 1988, *48*, 387–99.

Rosenberg, Nathan, *Perspectives on Technology*, New York: Cambridge University Press, 1976.

______, "Why in America?" in Otto Mayr and Robert Post, eds., *Yankee Enterprise: The Rise of the American System of Manufactures*. Washington: Smithsonian Institution Press, 1981.

______, "The Commercial Exploitation of Science by American Industry," in Kim B. Clark, Robert H. Hayes, and Christopher Lorenz, eds., *The Uneasy Alliance*, Boston: Harvard Business School Press, 1985.

Scott, Bruce R. and Lodge, George C., eds., *United States Competitiveness in the World Economy*. Boston: Harvard Business School Press, 1985.

Simon, Matthew and Novack, David E., "Some

Dimensions of the American Commercial Invasion of Europe, 1871–1914," *Journal of Economic History*, December 1964, *24*, 591–605.

Smith, George Otis, ed., *The Strategy of Minerals*. New York: D. Appleton and Company, 1919.

Spengler, Joseph J., ed., *Natural Resources and Economic Growth*. Washington: Resources for the Future, 1961.

Stern, Robert M. and Keith E. Maskus, "Determinants of the Structure of U.S. Foreign Trade, 1958–1976," *Journal of International Economics*, 1981, *11*, 207–24.

Vanek, Jaroslav, *The Natural Resource Content of United States Foreign Trade 1870–1955*. Cambridge, MA: MIT Press, 1963.

Vernon, Raymond, "International Investment and International Trade in the Product Cycle," *Quarterly Journal of Economics*, May 1966, *80*, 190–207.

______, ed., *The Technology Factor in International Trade*, New York: Columbia University Press, 1970.

Waehrer, Helen, "Wage Rates, Labor Skills, and U.S. Foreign Trade," in Peter Kenen and Roger Lawrence, eds., *The Open Economy*, New York: Columbia University Press, 1968.

White, Hal, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48*, 817–38.

Williamson, Oliver, "Emergence of the Visible Hand," in Alfred D. Chandler and Herman Daems, eds., *Managerial Hierarchies*. Cambridge, MA: Harvard University Press, 1980.

Woodruff, William, *America's Impact on the World*, New York: Wiley & Sons, 1975.

American Petroleum Institute, *Basic Petroleum Data Book*, September 1988, *8*.

Atlantic Institute, *The Technology Gap: The United States and Europe*. New York: Praeger, 1970.

International Geologic Congress, *The Iron-Ore Resources of the World*, Stockholm, 1910.

Markets of the United States, The Annals, September 1926, *127*.

United Nations, *Survey of World Iron-Ore Resources*, New York, 1955.

U.S. Department of Commerce, *Foreign Commerce and Navigation of the United States*, Washington: various years.

U.S. Geological Survey, *Mineral Resources of the United States*, Washington: 1913.

U.S. National Resources Committee, *Technological Trends and National Policy*. Washington: 1937.

U.S. Senate, *Report of the Immigration Commission*, Vol. 1, Washington: 1911.

Unexpected Inflation, Real Wages, and Employment Determination in Union Contracts

By DAVID CARD*

This paper examines the effect of nominal contracting provisions on employment determination in union contracts. In most contracts the nominal wage rate is wholly or partially predetermined. Real wage rates therefore contain unanticipated components that reflect unexpected price changes and the degree of indexation. The empirical analysis, based on a large sample of indexed and nonindexed contracts, suggests that unexpected real wage changes are associated with systematic employment responses in the opposite direction. I conclude that nominal contracting provisions play a potentially important role in the cyclical properties and persistence of employment movements in the union sector. (JEL 130, 820)

What role do nominal wage contracts play in the determination of employment and the characteristics of the business cycle? An influential series of papers by Stanley Fischer (1977), Edmund S. Phelps and John B. Taylor (1977), and John B. Taylor (1980) argued that fixed wage contracts create a link between employment and aggregate demand. More recent models of macro fluctuations stress other channels for the transmission and persistence of aggregate shocks. Real business cycle models (for example, Finn E. Kydland and Edward C. Prescott, 1982) assume that supply and demand in the labor market are equilibrated at Walrasian levels and ignore the institutional structure of wage determination. Recent models in the Keynesian tradition, on the other hand, have shifted attention from nominal wage rigidities to real wage rigidities (for example, Olivier J. Blanchard and Lawrence H. Summers, 1986) or nominal price rigidities (for example, N. Gregory Mankiw, 1985; Olivier J. Blanchard and Nobuhiro Kiyotaki, 1987).

This shift in interest reflects dissatisfaction with both the theoretical underpinnings and empirical performance of nominal contracting models. One the one hand, there are as yet no convincing theoretical explanations for the existence of nominally fixed contracts. Many of the models developed over the past decade predict constant real wages or constant real earnings.[1] On the other hand, the evidence in support of nominal contracting models is also weak. The simplest of these models asserts that aggregate demand shocks lead to real wage changes that induce movements along a downward-sloping demand schedule. Although unanticipated price increases are apparently correlated with real economic activity (see the review by Jo Anna Gray and David Spencer, forthcoming), the absence of a clear negative correlation between aggregate employment and real wages (Patrick T. Geary and John Kennan, 1982) poses a serious challenge to models of nominal wage rigidity.

This paper presents new evidence on the consequences of nominal contracting provisions for employment determination in the unionized sector of Canadian manufacturing. The analysis, based on data for 1300

*Department of Economics, Princton University, Princeton, NJ 08540. I am grateful to Robert Hall, Robert King, and two referees for their comments on earlier drafts. Thomas Lemieux and Sara Turner provided expert assistance in data preparation.

[1] See the survey of implicit contracting models by Sherwin Rosen (1985). A concise summary of the implications of these models from a macroeconomic perspective is presented by Stanley Fischer (1987, pp. 42–50).

indexed and non-indexed contracts written between 1966 and 1982, suggests that nominal contracting provisions play an important role in the link between aggregate demand and employment. As predicted by the simple models of Fischer (1977) and Jo Anna Gray (1976), I find that real wage changes induced by aggregate price surprises lead to systematic employment responses in the opposite direction. Unexpected real wage changes also affect subsequent wage determination: the empirical results suggest that roughly one-third of such changes carry over to the following contract. Unanticipated price increases therefore generate short-run employment responses and persistent wage changes among firms in the union sector.

Two features of the empirical analysis distinguish these results from earlier attempts to measure the effects of nominal wage rigidities. First, the analysis is based on individual contract data rather than aggregate or industry-level data.[2] Since union contracts differ in their negotiation dates and degrees of indexation, it is possible to calculate contract-specific measures of unexpected price increases and unexpected real wage changes, and to estimate the separate effects of price surprises and real wage surprises. Variation in contract lengths and the staggering of expiration dates also make it possible to control for aggregate-level disturbances that affect all contracts at a point in time. Second, the analysis pays special attention to the issue of endogenous wage determination.[3] Even in a simple Fischer-Gray contracting framework this is a potentially serious problem, insofar as the bargaining parties have information on future employment demand that is unavailable to an outside data analyst. If predictable components of future employment demand affect wages, they create a simultaneity bias in ordinary least-squares estimates of the elasticity of employment with respect to realized wage rates.

To solve this problem I use the unexpected component of real wages as an instrumental variable for the level of wages. By assumption, unexpected changes in real wages are correlated with wages but uncorrelated with information known at the negotiation date of the contract. Unexpected wage changes therefore form a valid instrumental variable for a structural analysis of employment demand. This procedure also provides a direct test of the role of nominal wage rigidities in generating employment responses to nominal shocks. The instrumental variables estimate of the elasticity of labor demand is nonzero if and only if employment is correlated with unexpected real wage changes.

The empirical results confirm the value of this approach. In ordinary least-squares regressions, changes in employment are only weakly related to changes in contract wages. When unexpected real wage changes are used as an instrumental variable, however, employment is found to be systematically negatively related to wages. This finding continues to hold when unexpected price changes are added directly to the employment demand equation. It is also robust to the addition of unrestricted dummy vari-

[2] Much of the earlier literature on nominal contracting models focuses on their implications for aggregate price and wage dynamics: see Taylor (1980) and Orley Ashenfelter and David Card (1982). A recent study by Shaghil Ahmed (1987) correlates the degree of wage flexibility in an industry, measured by the elasticity of indexation among indexed labor contracts, with the slope of the industry-specific Phillips curve. Ahmed's measure of wage flexibility is based on a sample of only 98 contracts in 20 industries, and fails to take into account any of the characteristics of the nonindexed contracts in an industry. Furthermore, his measure of flexibility only pertains to workers in large union contracts and ignores variation across industries in the extent of unionization or the share of large firms. Thus, I do not interpret his findings as strong evidence for or against the hypothesis that nominal contract rigidities are important. The approach taken by Mark Bils (1989) is perhaps most similar to that in this paper. He compares the variability of industry employment growth in months with a significant number of contract negotiations to the variability in other months.

[3] John Kennan (1988) presents an illuminating analysis of the difficulties that arise in the interpretation of aggregate employment and wage data when the data are generated by a simple model of demand and supply.

ables representing each year of the sample. I conclude that nominal wage contracts play an important role in determining the cyclical properties and persistence of employment in the union sector.

I. Employment and Wages in a Simple Contract Model

A. *Interpreting the Correlation of Employment and Wages*

This section outlines a simple model of long-term contracting in which nominal wages are predetermined and employment is set unilaterally by the firm after aggregate prices and firm-specific demand shocks are observed. Even in this simple model the interpretation of the partial correlation of employment and real wages is clouded by the fact that the contracting parties may have better information on future demand shocks than is available to an outside data analyst. To develop this point more formally, suppose that wages are negotiated in some base period (period 0) for a contract of duration T. Let $n(t)$ and $w(t)$ denote the logarithms of employment and real wages in period t of the contract, respectively, and assume that hours per worker are fixed. The notion of "nominal contracting" is captured by the assumption that the bargaining parties do not set $w(t)$ directly: rather, they establish a series of nominal wage increases from the start of the contract, possibly in conjunction with an indexation formula.[4] Let $w^*(t)$ represent the parties' expectation of $w(t)$, conditional on their information in the negotiating period, and let $u(t)$ represent the forecast error $w(t)-w^*(t)$. The distribution of $u(t)$ depends on the length of the contract and whether or not it contains a cost-of-living escalation clause.[5]

Assume that $n(t)$ is determined by an employment demand schedule of the form

$$(1) \quad n(t) = \alpha z(t) + \beta w(t) + \varepsilon(t),$$

where $z(t)$ is a vector of observable variables shifting the demand for labor, β represents the elasticity of labor demand ($\beta < 0$), and $\varepsilon(t)$ is an unobservable component of employment variation. The specification of $z(t)$ and the corresponding interpretation of β are discussed in the next section. Note that supply considerations are explicitly ignored: there are assumed to be enough available workers to fill the firm's demand irrespective of the forecast error in real wages. This assumption is a plausible one in the context of the available data, which pertain to unionized manufacturing establishments.

This simple model is completed by a specification of the determinants of $w^*(t)$. Assume that the expected real wage rate in period t is determined at the negotiation date by variables known at that time, say $x(0)$, and by the parties' expectations of $z(t)$ and $\varepsilon(t)$, $z^*(t)$ and $\varepsilon^*(t)$, respectively:

$$(2) \quad w^*(t) = az^*(t) + bx(0) + c\varepsilon^*(t).$$

The realized real wage rate in the tth period of the contract is therefore

$$w(t) = az^*(t) + bx(0) + c\varepsilon^*(t) + u(t).$$

The presence of simultaneity bias in ordinary least squares (OLS) estimates of the employment demand equation (1) depends on two factors. If $\varepsilon^*(t) \equiv 0$, then the parties have no informational advantage and there is no simultaneity problem. Alternatively, if $c = 0$, negotiated wages are unaffected by expected employment demand and again there is no simultaneity problem. If the parties are better able to forecast employment

[4] The nature of typical indexation formulas in North American labor contracts is described in my 1983 paper. The only case in which the real wage is set directly by the parties is the case of a contract in which nominal wages are indexed to the consumer price level with a formula that increases the wage by one percent for each percentage point increase in prices. Such formulas are rare, particularly in the manufacturing sector of the United States and Canada.

[5] This point is made by Wallace E. Hendricks and Lawrence M. Kahn (1987).

demand than an outside observer, however, and if higher forecasted demand leads to an increase in negotiated wage rates, then real wage rates will be positively correlated with the error in the employment equation, leading to a positively biased estimate of the wage elasticity β.

Irrespective of the parties' wage setting behavior, the elasticity β may be consistently estimated by considering the correlation between unanticipated wage rates and employment outcomes. The forecast error $u(t)$ forms a natural instrumental variable for $w(t)$: by definition, it is correlated with wages but uncorrelated with information available to the parties at the time of their negotiations.[6] Additional instruments may also be available if there are determinants of negotiated wages that can be legitimately excluded from the employment demand equation (the variables denoted as $x(0)$ in equation (2) above).

There are two important caveats to this procedure. The first is the possibility that forecast errors in real wages are directly correlated with unobservable determinants of labor demand. Suppose for example that employment demand shocks are positively correlated with unexpected price increases.[7] Then unexpected real wage increases are negatively correlated with employment demand shocks, leading to a negative bias in the instrumental variables estimate of the wage elasticity β. A simple way to control for this possibility is to include unexpected consumer price increases directly in the employment equation and to use variation across contracts in the degree of indexation to separately identify the effects of unexpected wage changes and unexpected price changes. A complementary approach is to include dummy variables representing the year in which employment is measured. These year effects absorb any aggregate demand shocks (or supply-side shocks) that affect all contracts in any given year.

A second difficulty may arise if unexpected changes in real wages during the term of a contract are immediately offset in subsequent negotiations. If this is the case then unexpected changes in real wages are inherently short-lived, and the presence of adjustment costs will substantially dampen the employment responses to such changes.[8] In the empirical analysis reported below I investigate the effect of real wage surprises on subsequent wage negotiations, and find that real wage rates in the subsequent contract move in the same direction as unexpected wage changes occurring during the previous contract. Thus, unexpected changes in real wages generate persistent effects on the cost of contractual labor, and should be expected to generate significant employment effects if the wage elasticity β is nonzero.

B. *Specification of the Employment Demand Function*

This section discusses the specification of the employment demand function (1) introduced above. An important limitation of the contract-based data set used in the empirical analysis is the absence of firm-specific price or output data. Selling prices, intermediate input prices, and output indexes are only available at the three-digit industry level. Nevertheless, these industry-level data may be used as proxies for the underlying firm-specific variables. To derive an interpretation of the resulting specification, suppose that output is produced from three

[6] It is interesting to compare this procedure to the one suggested by Bennett T. McCallum (1976) for the estimation of a structural equation that contains the expected value of a future endogenous variable. McCallum's procedure replaces the expected future value by its actual value and uses the predicted value (from a linear forecasting equation) as an instrumental variable. His procedure therefore eliminates simultaneity bias induced by a correlation between the dependent variable and the unexpected component of the explanatory variable. In the present context, the simultaneity bias arises from a correlation between the dependent variable and the expected value of the explanatory variable. Hence, the proposed instrument is the unexpected component of the explanatory variable.

[7] This may arise if employers have imperfect information on their relative demand shocks.

[8] Unexpectedly low real wage rates could induce an increase in overtime hours, however.

factors: labor, capital, and intermediate inputs (raw materials and energy). Ignoring firm-specific constants, assume that the logarithm of employment at a given firm in a particular industry in period $t, n(t)$ is related to the logarithm of firm-specific output, $y(t)$, the logarithm of firm-specific wages, $w(t)$, the logarithm of firm-specific nonlabor input prices, $v(t)$, the user cost of capital in period $t, r(t)$ (assumed to be constant across firms and industries), and an error term $\eta(t)$:

$$(3)\quad n(t) = \beta_1 w(t) + \beta_2 v(t) - (\beta_1 + \beta_2) r(t) + \sigma y(t) + \eta(t).$$

This equation can be derived from an underlying Cobb-Douglas production function, or alternatively it can be interpreted as a loglinear approximation to an arbitrary employment demand equation. The restriction that the elasticities of employment demand with respect to the three factor prices sum to zero is a consequence of the homogeneity of the cost function. This restriction implies that the equation is invariant to the deflator used to index wages and other factor prices. The magnitude of the coefficient σ reflects the degree of returns to scale: constant returns to scale implies $\sigma = 1$.

Let $\bar{y}(t)$ represent the logarithm of industry output in period t, and let $\bar{w}(t)$ and $\bar{v}(t)$ represent weighted averages of wages and intermediate input prices in the industry. Ignoring constants, assume that the logarithm of the firm's relative share of industry output is given by

$$(4)\quad y(t) - \bar{y}(t) = \gamma_1 (w(t) - \bar{w}(t)) + \gamma_2 (v(t) - \bar{v}(t)) + \phi(t).$$

This equation can be derived by assuming that firms with identical Cobb-Douglas production functions act as price takers with respect to firm-specific selling prices.[9] Alternatively, equation (4) can be interpreted as an approximation to the output share equation arising from a simple differentiated product oligopoly model. In either case, the error component $\phi(t)$ represents a mixture of firm-specific relative demand shocks and firm-specific productivity shocks.

The combination of equations (3) and (4) leads to an expression for firm-specific employment in terms of firm-specific wages, industry-level output and intermediate input prices, the aggregate cost of capital, and industry wages:

$$(5)\quad n(t) = (\beta_1 + \sigma\gamma_1) w(t) + \beta_2 \bar{v}(t) - (\beta_1 + \beta_2) r(t) + \sigma \bar{y}(t) - \sigma\gamma_1 \bar{w}(t) + (\beta_2 + \sigma\gamma_2)(v(t) - \bar{v}(t)) + \sigma\phi(t) + \eta(t).$$

Under the assumption that increases in marginal cost at a particular firm lead to decreases in its relative share of industry output, the coefficients γ_1 and γ_2 are negative. Thus, the elasticity of employment with respect to firm-specific wages, holding constant *industry* output, is larger in absolute value than the elasticity holding constant *firm-specific* output. Under the assumption of price-taking behavior the elasticity holding constant industry output is the unconditional elasticity of employment with respect to wages, allowing for the effect of changes

[9] Specifically, the Cobb-Douglas assumption implies that the output supply equation of the ith firm can be written as

$$y(t) = \gamma_1 w(t) + \gamma_2 v(t) + \gamma_3 r(t) - (\gamma_1 + \gamma_2 + \gamma_3) q(t) + \theta(t).$$

where $q(t)$ is the selling price for the output of the firm and $\theta(t)$ represents a total factor productivity shock. Define industry output as a geometric weighted average of the outputs of the individual firms in the industry. Then aggregate output follows a similar equation, and equation (4) can be derived directly, with

$$\phi(t) \equiv \theta(t) - \bar{\theta}(t) - (\gamma_1 + \gamma_2 + \gamma_3)(q(t) - \bar{q}(t)).$$

in wages on the output supply decision of the firm. Under these same assumptions the predicted elasticity of employment with respect to industry wages is positive, reflecting the fact that as industry wages increase (holding constant the firm's wage) the firm's share of industry output will increase.

C. *Allowing for the Presence of Efficient Contracting*

The specification of equation (5) assumes that employment levels are determined by the firm taking the realized real wage rate as given. Except under very special circumstances, however, unilateral employment determination by the firm fails to provide an efficient allocation of employment between contractual and extra-contractual opportunities.[10] For this reason, the empirical relevance of simple nominal contracting models has been sharply criticized (see Robert J. Barro, 1977, and Robert E. Hall, 1980, for example). The efficient determination of contractual employment is formally addressed in the implicit contracting literature and also the more recent efficient contracting literature.[11] The point of both literatures is that a jointly optimal contract (i.e., one that maximizes profit subject to a utility constraint for workers) determines employment on the basis of a shadow wage that can differ from the contractual wage. A contracting model with homogeneous workers and unrestricted transfers between employed and unemployed workers implies that the appropriate shadow wage is the marginal productivity of workers in their best alternative job. Brown and Ashenfelter (1986) refer to this as the "strong form" efficient contracting hypothesis. Strong form efficiency implies that contractual wages (and contractual wage rigidities) are irrelevant for employment determination and serve only to transfer income between employers and employees.[12]

In light of the differing implications of efficient contracting models and models with unilateral employment determination, it is important to adopt an empirical framework that encompasses either possibility. In principle this can be accomplished by including a measure of the appropriate shadow wage of labor in the employment demand function. A convenient assumption is that the shadow wage in an efficient contract is a weighted average of the observed contract wage and some measured alternative wage.[13] This leads to a specification of employment demand that includes both the contract wage and the measured alternative wage. Even though this procedure cannot provide a definitive test against the efficient contracting hypothesis,[14] it can provide useful evidence for or against the unilateral employment determination model, when the alternative is a testable version of the efficient contracting hypothesis.

II. Data Description and Measurement Framework

The empirical analysis in this paper is based on a sample of 1293 contracts negotiated by 280 firm and union bargaining pairs in the Canadian manufacturing sector.[15] The available information for each contract in-

[10] See Robert E. Hall and David Lilien (1979).

[11] The implicit contracts literature is reviewed by Sherwin Rosen (1985). See Ian M. McDonald and Robert M. Solow (1981) for a theoretical treatment of efficient contracting and James N. Brown and Orley Ashenfelter (1986) for a concise summary of the empirical implications of simple efficient contracting models.

[12] See John M. Abowd (1989) for an attempt to test this hypothesis using stock market data on negotiating firms.

[13] This hypothesis can be motivated formally by assuming that employees' preferences are represented by a Cobb-Douglas utility function defined over employment and the difference between the contractual wage and the alternative wage: see Brown and Ashenfelter (1986, p. S54).

[14] See Thomas E. MaCurdy and John H. Pencavel (1986), especially p. S13.

[15] The data set only includes contracts with 500 or more workers. The sample is drawn from a public use tape distributed by Labour Canada. A complete description of the sample and its derivation is presented in the Data Appendix. Louis N. Christofides and Andrew J. Oswald (1987) have also analyzed employment and wage data drawn from this source.

TABLE 1—CHARACTERISTICS OF EXPIRING CONTRACTS BY YEAR

Year	Number of Contracts (1)	Average Duration (2)	Percent with Escalation Clause (3)	Real Wage Index[a] 1971 = 100 (4)	Employment Index[b] 1971 = 100 (5)	Average Forecast Error[c] Prices (6)	Average Forecast Error[c] Real Wages (7)
1968	5	11.2	0.0	87.6	104.4	−0.1	0.1
1969	23	21.9	0.0	89.5	101.8	−0.9	0.9
1970	87	26.9	12.6	94.1	108.0	−2.0	1.8
1971	68	29.0	17.6	100.0	100.0	−4.6	3.8
1972	76	26.3	14.5	104.6	103.6	−3.0	2.8
1973	90	28.9	11.1	104.8	103.3	1.1	−1.1
1974	82	29.4	28.0	104.5	110.4	7.1	−6.1
1975	92	26.9	32.6	106.2	105.9	7.0	−6.3
1976	104	25.6	52.9	115.2	108.1	1.9	−1.2
1977	113	23.7	50.4	118.9	105.7	−2.2	1.8
1978	134	22.1	27.6	118.5	105.6	0.1	−0.3
1979	81	22.7	34.5	118.2	112.8	1.1	−0.9
1980	114	24.8	37.7	117.8	112.1	1.9	−1.2
1981	64	25.9	40.6	115.9	109.9	4.5	−3.3
1982	85	27.4	38.8	119.1	111.7	4.9	−3.8
1983	75	28.5	65.3	122.2	104.6	−0.5	1.2
Overall	1293	25.9	32.9	–	–	1.2	−0.9

Source: See Data Appendix.

[a] Estimated wage index for level of real wages at the end of expiring contracts.

[b] Estimated employment index for level of employment at the end of expiring contracts.

[c] Average percentage difference between price level (or real wage) at the end of contract and expected price level (or real wage) as forecast at the signing date of contract. See text.

cludes its starting (or effective) date, its ending (or expiration) date, and the base wage rate in each month of the contract.[16] The number of employees covered by the agreement is only available at renegotiation dates. I associate this level of employment with the expiring agreement. Thus, each sample point consists of an end-of-contract employment observation and a series of wages, including the beginning-of-contract and end-of-contract wage rates.

Some summary characteristics of the sample are presented in Table 1. The expiration dates of the contracts span a 16-year period between 1968 and 1983, with relatively few contracts in the first 2 years. The average duration of the contracts is 26 months, although durations vary somewhat by year, with relatively shorter contracts in the mid-1970s. The fraction of contracts with escalation clauses shows a steadily increasing trend until the mid-1970s and then varies erratically, with an overall average of 33 percent.

An indication of the trends in employment and wages in the sample is provided by the indexes in columns (4) and (5) of the table.[17] Real wage rates among expiring contracts show significant growth until 1977 and then remain relatively constant. Average employment shows no secular trend but reflects cyclical downturns in 1971, 1975, and 1983.

The empirical strategy of this paper is to fit regressions based on equation (5) to end-

[16] The base wage rate is typically the wage paid to the lowest-skill group covered by the collective bargaining agreement. An important assumption for the analysis in this paper is that variation over time in intracontract wage differentials is small enough to be safely ignored.

[17] The wage and employment indexes represent estimated year effects from regression equations for contract-to-contract percentage changes in end-of-contract wages and employment. These indexes therefore control for the composition of the set of expiring contracts in each year.

of-contract observations on employment and wages for each contract. Assuming that the employment demand function is homogeneous of degree zero in factor prices, the analysis is invariant to the choice of deflators for wages and intermediate input prices. Given the nature of wage indexation clauses, however, it is particularly convenient to work with real wages deflated by the consumer price index. In the remainder of the paper, wages and industry prices are therefore expressed as real variables, deflated by the consumer price index.

The real wage rate at the end of each contract is measured directly. This rate differs from its expectation as of the negotiation date of the contract by a component that depends on the indexation provisions of the contract and the deviation between actual and expected prices at the end of the contract. Following the notation above, let $w^*(T)$ represent the expected value of the logarithm of the real wage at the end of the contract. In a nonindexed contract, the logarithm of the actual real wage rate at the end of the contract, $w(T)$, is related to $w^*(T)$ by

$$(6)\quad w(T)=w^*(T)-(p(T)-p^*(T)),$$

where $p(T)$ represents the logarithm of the consumer price index at the end of the contract, and $p^*(T)$ represents the parties' expectation of $p(T)$, formed T months ago at the negotiation date of the contract.

In an indexed contract, unexpected changes in prices generate unexpected changes in real wage rates only to the extent that indexation is incomplete. For example, if an escalation clause increases nominal wages by e percent for each one percent increase in the consumer price index, then $w(T)$ and $w^*(T)$ are related by

$$(7)\quad w(T)=w^*(T)-(1-e)\times(p(T)-p^*(T)).$$

Although most escalation clauses in North American labor contracts do not specify a fixed elasticity of indexation, this equation is approximately correct when e is defined as the marginal elasticity of indexation evaluated at the expected level of prices at the end of the contract.

Given an estimate of the elasticity of indexation, $\hat{e}$, and an estimate of the parties' expected price level at the end of the contract, $\hat{p}(T)$, it is possible to decompose the real wage rate at the end of a contract into an estimate of its expected component, $\hat{w}(T)$, and an estimate of its unexpected component:

$$w(T)=\hat{w}(T)+\hat{u}(T),$$

where

$$\hat{w}(T)\equiv w(T)+(1-\hat{e})(p(T)-\hat{p}(T)).$$

Using the definition of $\hat{w}(T)$, the estimated unexpected component of real wages can be written as

$$\hat{u}(T)=u(T)+(\hat{e}-e)\times(p(T)-p^*(T))+(1-\hat{e})(\hat{p}(T)-p^*(T)).$$

This estimate differs from the true value $u(T)$ by two terms: one that depends on the difference between the actual and measured elasticity of indexation (and is therefore identically zero in a nonindexed contract), and another that depends on the difference between measured price expectations and the parties' true expectations. Provided that the measurement errors in the indexation elasticity and the expected price level are orthogonal to unmeasured components of employment demand, however, these errors do not preclude the use of $\hat{u}(T)$ as an instrumental variable for the level of wages at the end of the contract.

In this paper I use a naive forecasting model to form estimates of the expected price level at the end of the contract, based on the average rate of inflation over the 12 months prior to the negotiation date.[18] This

[18] The forecasting equation predicts the one-year ahead inflation rate at the negotiation date t as 0.0144 +0.7858 $DP(t-12)$, where $DP(t-12)$ is the actual

model was selected by comparing the noncontingent wage increases in the first year of 24–36 month nonindexed contracts to alternative forecasts of the 12-month inflation rate formed at the negotiation date of the contract. I have also experimented with more sophisticated forecasting equations and found few differences in the results. Since the forecasts are only used to form instrumental variables, the choice of an inefficient forecasting model should not bias the empirical results.

The other ingredient in the calculation of unexpected real wage changes is the elasticity of indexation e. Precise information on the actual indexation formulas in the sample is not readily available. I therefore use the ratio of total escalated increases over the life of the contract to the total increase in consumer prices over the life of the contract as a rough estimate of e. This measure is reasonably accurate for contracts with no restrictions on the escalation formula. For contracts with restricted escalation formulas that delay the start of indexation or specify a maximum escalated wage increase, this measure introduces some noise into the calculation of $\hat{u}(T)$.

Column 6 of Table 1 reports the average forecasting errors in the end-of-contract price level. The average annual forecast error is 1.2 percent, but it varies considerably by year, ranging from 7.0 percent for contracts expiring in 1974 and 1975, to -4.5 percent for contracts expiring in 1971. As the formulas in equations (6) and (7) imply, forecasting errors in end-of-contract real wage rates are negatively correlated with the forecast errors in prices. The average forecast errors in real wages in column 7 of the table are close to mirror images of the associated price forecasting errors. Relative to the forecasting errors in prices, however, the forecast errors in real wages are dampened by the indexation provisions of the escalated contracts. The average estimated elasticity of indexation among indexed contracts is 0.50, implying that the forecast errors in real wages among these contracts are about one-half as large as the corresponding forecast errors in prices.[19]

The average forecast errors in end-of-contract real wages are also negatively correlated with the employment index in column (5): the correlation coefficient over 16 annual observations is -0.54, and the regression coefficient of the employment index on unanticipated real wage changes is -0.70, with a standard error of 0.27. This provides some evidence that contractual employment outcomes are negatively related to unexpected changes in real wages. By comparison, the employment index is positively correlated with the index of real wage levels in column (4).

Contract-specific correlations between employment and wages are reported in Table 2. All the data in this table are measured as changes from the expiration date of the previous contract, using the sample of negotiations described in Table 1. Also presented in the table are the correlations of employment and wages with two measures of outside wages: the average real wage rate in the same (two-digit) industry, measured in the expiration month of the contract, and the average real wage for unskilled nonproduction laborers in the same province, measured in the expiration year of the contract.[20] Finally, the last two rows of Table 2 present the correlations of employment and wages with contract-specific measures of unexpected price changes and unexpected real wage changes.

percentage change in prices over the preceding 12 months. The two-and three-year-ahead inflation rate forecasts generated by this equation are $0.021+0.693$ $DP(t-12)$, and $0.026+0.6135$ $DP(t-12)$, respectively.

[19] The forecast error in end-of-contract real wages is $-(1-e)\rho$, where ρ is the forecast error in end-of-contract prices, and e is the elasticity of indexation. The average forecast error in real wages is therefore $-(1-\bar{e})\bar{\rho}+$ covariance(e,ρ), where $\bar{e}$ is the average elasticity of indexation and $\bar{\rho}$ is the average forecast error in prices.

[20] The provincial wage is measured from data collected annually by Labour Canada in its area wage survey. Data in this survey is collected by city. I have used the wage rate for the largest city in each province as a measure of the province-specific wage. See the Data Appendix.

TABLE 2—MEANS AND CORRELATION OF EMPLOYMENT AND WAGE CHANGES BETWEEN CONSECUTIVE CONTRACT[a]

	Mean	Standard Deviation	Employment (End of Contract)	Real Contract Wage (End of Contract)
1. Employment (End of Contract)	−0.017	0.201	1.00	−0.07
2. Real Contract Wage (End of Contract)	0.052	0.075	−0.07	1.00
3. Industry Wage (Expiration Month)	0.045	0.056	−0.04	0.59
4. Provincial Wage (Expiration Year)	0.044	0.060	−0.07	0.51
5. Unanticipated Change in Real Wages Over Contract[b]	−0.004	0.060	−0.12	0.45
6. Unanticipated Change in Consumer Prices Over Contract[c]	0.006	0.069	0.13	−0.44

[a] Sample size is 1293. All variables are measured as changes in logarithms between expiration dates of consecutive contracts.

[b] Percentage difference between real wage at end of contract and expected real wage forecast at signing date of contract.

[c] Percentage difference between Consumer Price Index at end of contract and expected price index forecast at signing date of contract.

These simple correlations reveal three features of the contract-level data. First, changes in employment are only weakly negatively correlated with changes in end-of-contract real wage rates. Second, the correlations between employment and outside wages are of similar magnitude to the correlations between employment and contract wages. Third, changes in employment are more strongly negatively correlated with changes in the unexpected component of real wages. Thus, the OLS estimate of the elasticity of employment with respect to contract wages is much smaller in absolute value than the corresponding instrumental variables estimate formed using unexpected changes in real wages as an instrumental variable. The OLS estimate is −0.19, with a standard error of 0.08, while the instrumental variables estimate is −0.70, with a standard error of 0.18. As will be seen below, this pattern continues to hold when other covariates are added to the employment determination equation.

III. The Effect of Previous Wage Rates on Subsequent Wage Determination

As a preliminary step in the analysis of employment demand, this section presents a brief summary of estimated wage equations for the sample of collective bargaining contracts described above. The purpose of this analysis is to identify any "spillover" effect from real wage rates at the end of one contract to wage rates in the next contract. A finding of significant spillovers implies that unexpected changes in real wages have persistent effects on the cost of contractual labor. A finding of insignificant spillovers, on the other hand, implies that these unexpected changes are relatively short-lived. The degree of persistence in unexpected wage changes is important for assessing the magnitude of the effect that these changes will exert on employment determination.

The analysis is based on two alternative measures of negotiated wages: the real wage at the start of the contract and the expected average real wage over the term of the entire contract. In the presence of adjustment costs the wage at the start of the next contract is particularly relevant for employment setting behavior in the last few months of an existing agreement. The expected average real wage over the next contract gives a longer-term measure of the costs of contractual employment.

A convenient statistical framework for analyzing the determinants of wages is a simple components-of-variance model of the form

$$(8)\quad w_{ij} = \theta_i + bx_{ij} + \lambda w(T)_{ij-1} + \xi_{ij},$$

where w_{ij} represents the measure of wages (either the real wage at the start of the contract or the expected average real wage over the life of the contract) for the jth contract of the ith firm, θ_i represents a permanent firm-specific component of wage variation, x_{ij} represents a vector of determinants of wages (measured at the negotiation date), $w(T)_{ij-1}$ represents the real wage at the end of the previous contract, and ξ_{ij} represents a contract-specific component of variance. The parameters b and λ can be estimated by taking contract-to-contract first-differences:

$$(9)\quad \begin{aligned} \Delta w_{ij} &\equiv w_{ij} - w_{ij-1} \\ &= b\Delta x_{ij} + \lambda \Delta w(T)_{ij-1} + \Delta\xi_{ij}. \end{aligned}$$

Ordinary least-squares estimates of this first-differenced wage equation may be inappropriate, however, if there is any correlation between the real wage at the end of the $(j-1)$st contract and the error component $\xi_{ij} - \xi_{ij-1}$ in the first-differenced wage equation.[21] This problem is readily overcome by using instrumental variables for the lagged change in ending real wage rates. Suitable instruments include the first-difference in the unexpected component of ending real wages and any exogenous components of Δx_{ij-1}. First-differencing also introduces a moving average error component into consecutive wage observations from the same bargaining pair. The estimated standard errors and test statistics throughout this paper therefore allow for a first-order moving average error component among the observations from each bargaining pair, as well as for arbitrary conditional heteroskedasticity.

Estimation results for the first-differenced wage equation (9) are reported in Table 3. Columns 1–4 of the table report estimates using the real wage at the start of the contract as the measure of wage outcomes, while columns 5–8 report estimates using the first-difference of the expected average real wage rate over the life of the contract as the dependent variable.[22] The components of x_{ij} include the regional unemployment rate and the real wage rate in aggregate manufacturing (measured in the effective month of the contract), a province-specific real wage rate for unskilled workers (measured in the effective year of the contract), and a set of unrestricted year effects for the effective date of the contract. The year effects capture a number of omitted factors, including a period of wage-price controls between 1975 and 1978. Their addition provides a significant improvement in the fit of the wage equations, although they hardly affect the estimated coefficient of previous wages. I have also estimated wage equations that include industry-specific output and price variables. These are only weakly related to negotiated wages, however, and their inclusion has virtually no effect on the reported coefficients in Table 3.

Colunms 1 and 5 of Table 3 report OLS estimates of equation (9) for the two alter-

[21] This problem is similar to one of estimating the effect of a lagged dependent variable in a panel data model: see Douglas Holtz-Aitken, Whitney Newey, and Harvey S. Rosen (1988).

[22] The expected average real wage in each month of the contract is estimated by formulas analogous to equations (6) and (7), using estimates of the expected price level in that month and estimates of the elasticity of indexation as described above. The expected average real wage is an unweighted average of expected monthly rates sampled at six-month intervals throughout the contract period, starting in the first month of the contract.

Table 3—Estimated Wage Determination Equations

	Real Wage at Start of Contract				Expected Average Real Wage During Contract			
	OLS	IV[a]			OLS	IV[a]		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Year Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
2. Regional Unemployment Rate	−0.50	−0.45	−0.46	−0.46	−0.38	−0.44	−0.45	−0.47
	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)	(0.13)	(0.13)	(0.13)
3. Real Wage in Manufacturing	0.04	0.11	0.11	0.10	0.40	0.30	0.31	0.26
	(0.10)	(0.11)	(0.11)	(0.12)	(0.11)	(0.12)	(0.12)	(0.12)
5. Real Wage in Region	0.02	0.04	0.04	0.03	0.04	0.02	0.02	0.01
	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
5. Real Wage at End of Previous Contract	0.48	0.36	0.35	–	0.25	0.41	0.35	–
	(0.03)	(0.05)	(0.07)		(0.03)	(0.06)	(0.07)	
6. Expected Real Wage at End of Previous Contract	–	–	–	0.46	–	–	–	0.36
				(0.08)				(0.09)
7. Unexpected Real Wage at End of Previous Contract	–	–	–	0.41	–	–	–	0.43
				(0.06)				(0.07)
8. Change in Prices During Previous Contract	–	–	−0.01	–	–	–	−0.05	–
			(0.03)				(0.03)	
9. Standard Error	0.039	0.039	0.039	0.038	0.038	0.039	0.038	0.038
10. Overidentification Test[b]	–	0.261	0.273	0.489	–	0.037	0.016	0.006

Note: Standard errors in parentheses. Sample size is 1293. All regressions include a (first-differenced) linear trend. The mean and standard deviation of the dependent variable in columns (1)–(4) are 0.050 and 0.066. The mean and standard deviation of the dependent variable in columns (5)–(8) are 0.043 and 0.061. Standard errors are corrected for first-order moving average error component and heteroskedasticity.

[a] In columns (2), (3), (6), and (7), instrumental variables for real wage at the end of the previous contract include 18-year effects, the real wage in manufacturing at the start of the previous contract and the unanticipated change in real wages over the previous contract. In columns (4) and (8) instrumental variables for expected real wage at the end of the previous contract include 18-year effects, the real wage in manufacturing at the start of the previous contract, and the change in consumer prices during the previous contract.

[b] Probability value of test for orthogonality of residuals and instruments. The statistic is distributed as chi-squared with 19 degrees of freedom in columns (2), (3), (6), and (7), and with 18 degrees of freedom in columns (4) and (8).

native dependent variables, while columns 2 and 6 report instrumental variables (IV) estimates. These specifications suggest that negotiated wages are significantly positively related to the level of wages at the end of the preceding contract. The OLS estimates of the spillover coefficient λ (in row 6) differ somewhat between the two alternative measures of the dependent variable, although the IV estimates are closer together. The last row of the table reports overidentification test statistics for the instrumental variables estimators. There is no evidence against the exclusion restrictions implicit in the IV procedure for the specification in column 2. The test statistic for the specification in column 6, on the other hand, presents mild evidence against these restrictions.

In columns 3 and 7 the change in prices over the preceding contract is introduced directly into the wage determination equation. This addition permits a test of the hypothesis that aggregate price movements affect future wage determination only to the extent that they affect the level of real wages at the end of the preceding contract. The estimated coefficients in row 8 of the table provide no evidence against this hypothesis. Finally, the specifications in columns 4 and 8 relax the assumption that the expected and unexpected components of the end-of-contract wage $w(T)_{ij-1}$ have the same effect on subsequent wages.[23] Perhaps sur-

[23] These equations are estimated using the change in prices over the previous contract, the manufacturing

prisingly, there is no evidence against the restricted specification: the t-statistics for the hypothesis of equal coefficients for the expected and unexpected components are 1.32 in column 4 and 1.22 in column 8.

These results suggest that unexpected changes in wages have persistent effects on the costs of contractual labor. An unanticipated 10 percent decrease in real wages leads to an approximately 3 percent lower real wage throughout the following contract. Thus even in the presence of substantial adjustment costs, employment should be expected to respond to unanticipated changes in real wages, provided that the unilateral employment determination model is correct.

IV. The Determinants of Contractual Employment

This section turns to estimates of the employment demand function (5). As in the previous section, the framework for the analysis is a components-of-variance model for the logarithm of end-of-contract employment in the jth contract of the ith firm (n_{ij}):

$$(10)\quad n_{ij} = \psi_i + \alpha z_{ij} + \beta w(T)_{ij} + \varepsilon_{ij}.$$

In this equation, ψ_i represents a permanent firm-specific effect, z_{ij} represents a vector of determinants of employment, measured at the end of the contract, $w_{ij}(T)$ represents the real wage rate at the end of the contract, and ε_{ij} is a contract-specific disturbance. Assuming that industry output and prices are used as proxies for firm-specific output and price data, the wage elasticity β in equation (10) is related to the underlying parameters of the employment demand schedule (3) and the relative output equation (4) by $\beta = -(\beta_1 + \sigma\gamma_1)$. Note that β is assumed to be constant across industries. Although this is unlikely to be true, the relatively small number of contracts in each industry makes it difficult to estimate parameters other than the average demand elasticity across industries. Heteroskedasticity introduced by variation in β is taken into account in the calculation of the standard errors.

Again, a convenient method for eliminating the pair-specific effects is to take first-differences between consecutive contracts, yielding

$$(11)\quad \Delta n_{ij} = \alpha \Delta z_{ij} + \beta \Delta w(T)_{ij} + \Delta \varepsilon_{ij}.$$

In many previous studies, employment outcomes have been found to follow a partial adjustment equation of the form $n_{ij} = (1-\mu)n^*_{ij} + \mu n_{ij-1}$, where n^*_{ij} represents the optimal level of employment in the absence of adjustment costs, as given by an equation such as (5). Partial adjustment is readily accommodated within the framework of equation (11) by the addition of a lagged dependent variable. In the present context, however, consecutive employment outcomes are 20–36 months apart. Thus, the extent of partial adjustment is likely to be much smaller than that observed in quarterly or annual data. This issue is addressed more thoroughly below.

Estimation results for the first-differenced employment equation are presented in Tables 4 and 5. Following the discussion in Section I, Part B, the determinants of employment include the three-digit industry input price index (deflated by the consumer price index), industry-level real output, and the end-of-contract real wage rate. Specifications that add outside wage rates and a lagged dependent variable are presented in Table 5. The odd-numbered columns of Table 4 present estimated equations that include a linear time trend, while the even-numbered columns report estimates that include a set of unrestricted dummy variables for the different expiration years in the sample. I have not made any attempt to measure the user cost of capital. On the assumption that capital costs are constant across manufacturing industries, variation in

wage at the effective date of the previous contract, and year effects for the effective date of the previous contract as instrumental variables for the expected and unexpected components of real wages at the end of the previous contract.

TABLE 4—ESTIMATED EMPLOYMENT DETERMINATION EQUATIONS

	OLS		IV^a		IV^b			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Year Effects	No	Yes	No	Yes	No	Yes	No	Yes
2. Real Industry Input Price	0.22	0.16	0.20	0.16	0.19	0.16	0.19	0.15
	(0.06)	(0.08)	(0.06)	(0.08)	(0.06)	(0.08)	(0.06)	(0.08)
3. Real Industry Output	0.20	0.29	0.22	0.28	0.23	0.28	0.23	0.28
	(0.07)	(0.09)	(0.07)	(0.09)	(0.07)	(0.09)	(0.07)	(0.09)
4. Real Industry Output	0.17	0.10	0.15	0.11	0.14	0.11	0.14	0.10
(Previous Year)	(0.06)	(0.07)	(0.07)	(0.07)	(0.06)	(0.07)	(0.07)	(0.07)
5. Real Wage at End of	−0.15	−0.02	−0.28	−0.45	−0.39	−0.51	−0.42	−0.40
Contract	(0.08)	(0.10)	(0.17)	(0.35)	(0.12)	(0.29)	(0.17)	(0.42)
6. Unexpected Inflation	–	–	–	–	–	–	−0.03	0.10
During Contract							(0.13)	(0.20)
7. Standard Error	0.196	0.194	0.196	0.195	0.196	0.196	0.196	0.195
8. Test for Exclusion of	–	0.003	–	0.006	–	0.004	–	0.004
Year Effects (p-Value)								
9. Overidentification	–	–	–	–	0.76	0.97	0.74	0.96
Test^c								

Note: Standard errors in parentheses. Sample size is 1293. All regressions include a (first-differenced) linear trend. The mean and standard deviation of the dependent variable are −0.017 and 0.201. Standard errors are corrected for first-order moving average error component and heteroskedasticity.

[a] Instrumental variable for real wage at end of contract is the unanticipated change in real wages during the contract.

[b] Instrumental variables for real wage at end of the contract include 18 year effects, the real wage in manufacturing at the start of the contract, and the unanticipated change in real wages during the contract.

[c] Probability value of test for orthogonality of residuals and instruments. The test statistic is distributed as chi-squared with 19 degrees of freedom in all cases.

the user cost of capital is absorbed by the trends and/or time effects in the empirical specification. The unrestricted year effects also capture any aggregate-level shocks (such as aggregate demand shocks or productivity shocks) that are shared by all contracts in a given year.

In an effort to capture partial adjustment effects, and also to control for the fact that industry output is measured annually, the employment equations in Tables 4 and 5 include industry output in both the expiration year of the agreement and the previous year. I have experimented with specifications that also include wage rates and input prices in the year prior to the expiration date, but the effects of these variables are always poorly determined and small in magnitude.

The first two columns of Table 4 present OLS estimates of the employment equation with and without dummy variables for the expiration date of the contract. Employment is positively related to intermediate input prices and current and last year's level of output. The elasticity of employment with respect to output (i.e., the sum of the coefficients of current and last years' output) is substantially less than unity, implying increasing returns to scale in the framework of equation (5). The addition of the year effects results in a relatively small improvement in the fit of the employment equations: the probablity value of an exclusion tests for the year effects is reported in row 8 of the table. When the year effects are included, however, the estimated wage elasticity of employment demand falls to essentially zero.

The estimated wage elasticity is substantially larger (in absolute value) when the end-of-contract wage rate is instrumented by the unanticipated change in real wages over the term of the contract. The results of this exercise are reported in columns 3 and 4 of Table 4. Without year effects, the esti-

TABLE 5—ESTIMATED EMPLOYMENT DETERMINATION EQUATIONS

	OLS		IV[a]			IV[b]	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. Year Effects	Yes	Yes	Yes	Yes	Yes	No	Yes
2. Real Industry Input Price	0.16	0.16	0.14	0.16	0.14	0.13	0.10
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.09)
3. Real Industry Output	0.29	0.29	0.27	0.28	0.27	0.20	0.25
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.07)	(0.09)
4. Real Industry Output (Previous Year)	0.10	0.10	0.11	0.11	0.11	0.15	0.13
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)
5. Real Wage at End of Contract	−0.03	−0.02	−0.56	−0.51	−0.56	−0.52	−0.58
	(0.10)	(0.10)	(0.32)	(0.31)	(0.33)	(0.22)	(0.32)
6. Real Wage in Industry	0.06	–	0.23	–	0.23	0.26	0.38
	(0.22)		(0.26)		(0.26)	(0.22)	(0.25)
7. Real Wage in Region	–	−0.03	–	0.04	0.06	–	–
		(0.15)		(0.16)	(0.21)		
8. Lagged Dependent Variable (Instrumented)	–	–	–	–	–	−0.13	−0.08
						(0.14)	(0.15)
9. Standard Error	0.194	0.194	0.196	0.196	0.196	0.193	0.194
10. Overidentification Test[c]	–	–	0.972	0.967	0.972	0.451	0.666

Note: See note to Table 4. Standard errors in parentheses.

[a] Instrumental variables for the real wage at the end of the contract include 18-year effects, the real wage in manufacturing at the start of the contract, and the unanticipated change in real wages during the contract.

[b] Estimated on subsample of 1107 observations. Mean and standard deviation of the dependent variable are −0.015 and 0.0200, respectively. Instruments include the instrument set above plus the lagged value of industry output.

[c] Probability value of test for orthogonality of residuals and instruments. The test statistic is distributed as chi-squared with 19 degrees of freedom in columns (3)–(5), and 16 degrees of freedom in columns (6)–(7).

mated elasticity rises from −0.15 to −0.28, although the estimated standard error rises proportionately. With year effects, the change in the point estimate is even more remarkable: from −0.02 to −0.45. Due to the imprecision of the IV estimators, however, tests of the difference between the OLS and IV estimates are insignificant in either case.

The specifications in columns 5 and 6 attempt to reduce this imprecision by expanding the list of instrumental variables for the end-of-contract real wage rate to include the level of real wages in manufacturing at the start of the contract and year effects for the signing date of the contract. The additional instrumental variables lead to a slight increase in the magnitude of the estimated wage elasticities and provide some increase in the precision of the estimates. Overidentification test statistics for the internal consistency of the instruments are reported in row 9 of the table. In all cases these are below conventional significance levels. I have also estimated employment equations that use only the additional instruments (i.e., excluding the unexpected change in real wages) to identify the effect of wages on employment. As the overidentification statistics suggest, these estimates are very similar to those in Table 4.

Even with the additional instrumental variables the estimated elasticity of employment demand in column 6 is only significantly different from zero at the 10 percent level. Nevertheless, a test of the difference between the estimated demand elasticities in columns 1 and 5 is significant at the 1 percent level, and a test of the difference between the estimated elasticities in columns 2 and 6 is significant at the 10 percent level. These results suggest that OLS estimates of the elasticity of employment demand are positively biased.

The final two columns of Table 4 present employment equations that include the unexpected change in consumer prices during the term of the contract as an additional explanatory variable. These specifications provide a simple check on whether unexpected price increases affect employment through the contractual wage, or whether there is a direct correlation between unexpected inflation and employment demand.[24] Neither specification provides any evidence of a direct role for unexpected price changes. Nevertheless, the standard errors of the wage and price terms in column 8 are sufficiently large that one cannot rule out a direct effect of inflationary surprises on employment demand.[25] Taken together with the other estimates in the table, however, I interpret the results in columns 7 and 8 as supporting the conclusion that price surprises affect employment determination solely through their effect on realized wages.

The effect of outside wage rates on contractual employment is addressed in Table 5. The theoretical analysis in Section I identifies two alternative routes for this effect. On one hand, increases in average wages in the industry may have a positive effect on employment, reflecting the competitive advantage implied by higher costs elsewhere in the industry. On the other hand, increases in wage rates representing the alternative value of workers' time may have a negative effect if employment is influenced by efficient contracting considerations. In an effort to distinguish between these hypotheses, I have included the industry average wage in columns 1 and 3 of the table, and a province-specific wage for unskilled laborers in columns 2 and 4 of the table. Both wage measures are included in column 5.

The OLS estimates in columns 1 and 2 of Table 5 show no evidence of a role for either outside wage measure. When the contract wage is instrumented, however, the point estimate of the effect of the industry-specific wage rises substantially, while the estimated effect of the regional wage measure remains close to zero. A similar pattern emerges in column 5 when both outside wage measures are included. Given the imprecision of the estimated elasticities it is difficult to draw strong conclusions from these results. Nevertheless, the estimates lend much stronger support to the view that outside wages belong in the employment equation as a proxy for the level of competitors' costs than to the view that outside wages belong in the employment equation as a proxy for the shadow value of employees' time.[26] If the former view is taken literally, the point estimates in column 3 suggest that the output-constant elasticity of employment demand with respect to wages is -0.33, while the elasticity of output supply with respect to an increase in wages is -0.70.[27] This estimate of the output-constant demand elasticity is in the midpoint of the range of estimates usually reported in the static employment demand literature (see Daniel S. Hamermesh, 1986, pp. 451–54).

The question of whether the estimated employment equations are robust to the inclusion of lagged employment is explored in the last two columns of Table 5. Since the employment models are estimated in first-differences, and the covariance of consecutive changes in employment is biased downward by any measurement error, the lagged value of industry output is added to the list of instrumental variables, and lagged employment and real wages are treated as jointly endogenous. The results show no evi-

[24] It is worth pointing out, however, that aggregate demand shocks (or any other variables that affect all contracts at a point in time) are absorbed by the year effects included in columns 4 and 6.

[25] At the suggestion of a referee, I estimated an employment equation that includes unexpected price increases (and year effects) and excludes wages. In this specification the estimated elasticity of employment with respect to unanticipated price increases is 0.23, with a standard error of 0.14.

[26] My 1986 paper and Stephen J. Nickell and Sushil Wadhwani (1987) report employment specifications that show a positive effect of outside wages, while Brown and Ashenfelter (1986) report positive effects in more than one-half of their specifications.

[27] Recall from equation (5) that the elasticity of employment with respect to wages is $-(\beta_1+\sigma\gamma_1)$, while the elasticity of employment with respect to industry average wages is $\sigma\gamma_1$. An estimate of σ from column (3) of Table 5 is 0.39 (the sum of the coefficients of current and last year's output). Using the other estimated coefficients from this equation leads to the estimates of β_1 and γ_1 reported in the text.

dence of a role for lagged employment. As mentioned earlier, this probably reflects the 20–36 month interval between consecutive observations in the data set. Over two or three years the effects of partial adjustment are likely to be much smaller than over an interval of a quarter or year.[28]

The estimates in Tables 4 and 5 suggest two main conclusions. First, employment outcomes are negatively related to contractual wage rates. Although the simple correlation between end-of-contract wage rates and employment is small and statistically insignificant, this is apparently a consequence of simultaneity bias. When unanticipated real wage changes and/or other exogenous variables are used as instrumental variables for the end-of-contract wage, the estimated wage elasticity is consistently negative and stable in magnitude across alternative specifications. Second, there is no evidence that employment is related to outside wages in a manner consistent with simple efficient contracting models. Even though employment is uncorrelated with region-specific wage measures, it is weakly positively correlated with industry average wages. This positive correlation is consistent with the hypothesis that higher average industry wages lead to improvements in the firm's competitive position and increases in employment.

V. Conclusions

This paper presents new evidence on the role of nominal wage contracts in the union sector. An important feature of these contracts, emphasized by the simple macro models of Fischer (1977) and Taylor (1980), is the predetermined nature of nominal wages. Real wage rates at the end of a contract therefore contain unanticipated components that reflect unexpected changes in consumer prices and the degree of indexation in the contract. The empirical analysis, based on a large sample of indexed and nonindexed contracts, indicates that these unexpected real wage changes are associated with systematic employment responses in the opposite direction. This suggests that nominal contracts play a role in the link between aggregate demand shocks and real economic activity, at least in the part of the economy covered by explicit nominal contracts.

Three other findings emerge from the empirical analysis. First, the contract-level correlation between employment and wages apparently reflects both demand and wage-setting behavior. Similar simultaneity problems may arise in other studies of firm-specific employment and wage data. Second, unanticipated changes in prices are found to generate changes in real wages that spill over from existing labor contracts to subsequent agreements. Inflation surprises therefore have persistent effects on real wages in the union sector, in addition to their short-run effects on employment. Finally, the empirical results suggest that employment outcomes in union contracts are determined on a conventional downward-sloping demand schedule, taking the prevailing contract wage as given. There is no indication that employment is related to outside wages in a manner consistent with a simple model of efficient contracting.

DATA APPENDIX

I. Contract Sample

The contract sample is derived from the December 1985 version of Labour Canada's Wage Tape. This tape contains information on collective bargaining agreements covering more than 500 employees in Canada. Starting from the 2868 manufacturing contracts on the tape, I merged together contract chronologies between the same firm and union covering different establishments, and eliminated contracts from bargaining pairs with fewer than four contracts. These procedures yield a sample of 2258 contracts negotiated by 299 firm and union pairs. Further information on the merging process and the characteristics of the resulting sample are presented in the Data Appendix to my 1988 paper and in Tables 1 and 2 of that paper.

The employment data for this sample were then checked in two stages. First, the number of workers covered in each contract was compared to the number covered in the preceding and subsequent agreements. Second, in cases where the number of workers changed dramatically between contracts, the contract summaries in the appropriate issue of the *Collective Bar-*

[28] In principle, the coefficient of the lagged dependent variable will differ, depending on the duration of the previous contract. In view of the imprecision of the estimated partial adjustment coefficients in Table 5, however, I have not attempted to address this issue.

gaining Review were consulted. In 238 contracts, the emplcyment counts recorded on the wage tape were found to be in disagreement with the counts reported in the *Collective Bargaining Review*. In these cases, counts from the published contract summaries were used. In cases for which the set of establishments covered by the contract changed over time, contracts with inconsistent coverage were deleted from the sample. Of the 2258 contracts in the subsample of merged contracts, valid coverage data are available for 1813 contracts (80.3 percent). Checking of the employment data was performed by Thomas Lemieux. I am extremely grateful for his assistance with these data.

In this paper, employment at the end of a contract is measured by the number of workers covered by the subsequent agreement. Furthermore, the estimation procedures require information on employment and wage outcomes in the previous agreement and on various industry and aggregate data that are only available between 1966 and 1983. The sample of contracts used in this paper therefore consists of the subset of contracts in the initial 2258 contract merged subsample that satisfy the following criteria:

(a) Information on at least one previous contract is available in the sample.

(b) Information on at least one subsequent contract is available in the sample.

(c) The expiration dates of the current and previous contract are after January 1966 and before December 1983.

(d) Valid employment data are available for both the current and preceding contract (i.e., valid counts of workers covered are available for both the current and subsequent contracts).

II. Aggregate and Industry-Level Data

The following aggregate and industry-level data were merged to the contract sample.

(a) Consumer price index, all items, 1981 = 100. January 1961 to November 1985: Cansim D484000, from the 1985 Cansim University Base Tape. December 1985 to June 1986: from the *Bank of Canada Review*, November 1986.

(b) Average hourly earnings in manufacturing. January 1961 to March 1983: Cansim D1518, from the 1983 Cansim University Base Tape. April 1983 to December 1983: Cansim L5607, from the *Bank of Canada Review*, various issues. Data from April 1983 and later are multiplied by 1.04035 to correct for the revision in the establishment survey.

(c) Average hourly earnings of nonproduction production laborers, by province. Annual data on hourly earnings for selected occupations are available for major cities. I matched data for the following cities to their respective provinces: Halifax, St. John, Montreal, Toronto, Winnipeg, Regina, Edmonton, Vancouver. The wage rates used are listed as rates for "male general laborers" between 1966 and 1977, for "general laborers in service occupations" between 1978 and 1981, and for "nonproduction laborers" between 1982 and 1985. Data for 1966–72 are from *Wage Rates, Salaries, and Hours of Labour*, 1966–1972 editions. Data for 1973–1986 are from *Canada Year Book*, various editions. For contracts that cover two or more provinces, I used a weighted average of Montreal, Toronto, and Vancouver rates with weights of 0.35, 0.55, and 0.10, respectively.

(d) Unemployment rates, seasonally adjusted. For contracts in Quebec, Ontario, and British Columbia, I used the province-specific unemployment rates for all workers. For contracts in other provinces, I used the national average unemployment rate. The series used were as follows: Quebec–Cansim D768478; Ontario–Cansim D768648; British Columbia–Cansim D769233; all others–Cansim D767611. Data for January 1966 through November 1983 were obtained from the 1983 Cansim University Base. Data for December 1983 were taken from the *Bank of Canada Review*, November 1986.

(e) Industry selling prices, input prices, and output. Three-digit industry level annual data for 1961–71 were taken from Statistics Canada, *Real Domestic Product by Industry 1961–71*. These data are classified by 1960 standard industrial codes (SICs). Data on a 1971 SIC basis for 1971–83 were taken from the 1978 and 1984 issues of Statistics Canada, *Gross Domestic Product by Industry*. The 1960 and 1971 SIC codes were then matched, and the price and output indexes spliced using the 1971 observations from the two sources. Of 65 three-digit industries represented in the contract sample, there were a total of 31 for which three-digit-level data were not available on a consistent basis. For these industries, two-digit-level data were used. The publications report the value of gross output and implicit price indexes for gross output and intermediate inputs. These data were used to construct the value of real gross output (the measure of "output" used in this paper). Implicit price indexes for gross output and intermediate inputs were deflated by the annual average consumer price index to obtain real selling prices and input prices used in the paper.

(f) Industry average hourly earnings. Monthly two-digit industry-level average hourly earnings data for the period January 1961 to March 1983 were taken from the 1983 Cansim University Base. Earnings data are unavailable for two industries: knitting mills and miscellaneous manufacturing. For the former, I used earnings in clothing industries. For the latter, I used average earnings in all manufacturing. Wage rates for April through December 1983 were constructed by index-linking wage rates from the new establishment survey to the rates in the old survey using their values in March 1983. Earnings data from the new survey for March-December 1983 were taken from the 1985 Cansim University Base.

REFERENCES

Abowd, John M., "The Effect of Wage Bargains on the Stock Market Value of the Firm," *American Economic Review*, September 1989, *79*, 774–800.

Ahmed, Shaghil, "Wage Stickiness and the

Non-Neutrality of Money: A Cross-Industry Analysis, " *Journal of Monetary Economics*, July 1987, *20*, 25–50.

Ashenfelter, Orley and Card, David, "Time-Series Representations of Economic Variables and Alternative Models of the Labour Market," *Review of Economic Studies*, Supplement 1982, *49*, 761–82.

Barro, Robert J., "Long Term Contracts, Sticky Prices, and Monetary Policy," *Journal of Monetary Economics*, July 1977, *3*, 305–16.

Bils, Mark, "Testing for Contracting Effects on Employment," National Bureau of Economic Research Working Paper No. 3051, July 1989.

Blanchard, Olivier J. and Summers, Lawrence H., "Hysteresis and the European Unemployment Problem," in Stanley Fischer, ed., *NBER Macroeconomics Annual*, Cambridge: MIT Press, 1986, 15–78.

______ **and Kiyotaki, Nobuhiro,** "Monopolistic Competition and the Effects of Aggregate Demand, " *American Economic Review*, September 1987, *77*, 647–66.

Brown, James N. and Ashenfelter, Orley, "Testing the Efficiency of Employment Contracts," *Journal of Political Economy*, June 1986, *94*, S40–S87.

Card, David, "Cost of Living Escalators in Major Union Contracts," *Industrial and Labor Relations Review*, October 1983, *37*, 34–48.

______, "Efficient Contracts with Costly Adjustment: Short Run Employment Determination for Airline Mechanics," *American Economic Reivew*, December 1986, *76*, 1045–71.

______, "Strikes and Wages: A Test of a Signalling Model," National Bureau of Economic Research Working Paper No. 2550, April 1988.

Christofides, Louis N. And Oswald, Andrew J., "Efficient and Inefficient Employment Outcomes: A Study Based on Canadian Contract Data," Oxford Institute of Economics and Statistics Applied Economics Discussion Paper No. 37, November 1987.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–206.

______, "Recent Developments in Macroeconomics," National Bureau of Economic Research Working Paper No. 2473, December 1987.

Geary, Patrick T. and Kennan, John, "The Employment–Real Wage Relationship: An International Study," *Journal of Political Economy*, August 1982, *90*, 854–71.

Gray, Jo Anna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, *2*, 221–35.

______ **and Spencer, David,** "Price Prediction Errors and Real Economic Activity: A Reassessment," *Economic Inquiry*, forthcoming.

Hall, Robert E., "Employment Fluctuations and Wage Rigidity," *Brookings Papers on Economic Activity*, 1980, 91–123.

______ **and Lilien, David,** "Efficient Wage Bargains Under Uncertain Supply and Demand," *American Economic Review*, December 1979, *69*, 868–79.

Hamermesh, Daniel S., "The Demand for Labor in the Long Run," in Orley Ashenfelter and Richard Layard, eds., *Handbook of Labor Economics*, Amsterdam: North-Holland, 1986, ch. 8.

Hendricks, Wallace E. and Kahn, Lawrence M., "Contract Length, Wage Indexation, and *Ex Ante* Variability of Real Wages," *Journal of Labor Research*, Summer 1987, *8*, 221–36.

Holtz-Eakin, Douglas, Newey, Whitney and Rosen, Harvey S., "Estimating Vector Autoregressions with Panel Data," *Econometrica*, November 1988, *56*, 1371–98.

Kennan, John, "An Econometric Analysis of Fluctuations in Aggregate Labor Supply and Demand," *Econometrica*, March 1988, *56*, 317–34.

Kydland, Finn E. and Prescott, Edeard C., "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

MaCurdy, Thomas E. and Pencavel, John H., "Testing Between Competing Models of Wage and Employment Determination in Unionized Markets," *Journal of Political Economy*, June 1986, *94*, S3-S39.

Mankiw, N. Gregory, "Small Menu Costs and Large Business Cycles: A Macroeconomic

Model of Monopoly," *Quarterly Journal of Economics*, May 1985, *100*, 529–39.

McCallum, Bennett T., "Rational Expectations and the Natural Rate Hypothesis: Some Consistent Estimates," *Econometrica*, January 1976, *44*, 43–52.

McDonald, Ian M. and Solow, Robert M., "Wage Bargaining and Employment," *American Economic Review*, December 1981, *71*, 896–908.

Nickell, Stephen J. and Wadhwani, Sushil, "Financial Factors, Efficiency Wages and Employment: Investigations Using U.K. Micro-Data," London School of Economics Center for Labour Economics Working Paper No. 993, November 1987.

Phelps, Edmund S. and Taylor, John B., "Stabilization Powers of Monetary Policy Under Rational Expectations," *Journal of Political Economy*, February 1977, *85*, 163–90.

Rosen, Sherwin, "Implicit Contracts: A Survey," *Journal of Economic Literature*, September 1985, *23*, 1144–76.

Bank of Canada Review, various issues, Ottawa: Bank of Canada.

Canada Department of Labour, *Wage Rates, Salaries and Hours of Labour*, various issues, Ottawa: Information Canada.

Labour Canada, *Collective Bargaining Review*, various issues, Ottawa: Labour Canada.

Statistics Canada, Information Division, *Canada Year Book*, various issues, Ottawa: Statistics Canada.

Statistics Canada, Industry Product Division, *Gross Domestic Product by Industry*, Ottawa: Statistics Canada, September 1984.

______, *Real Domestic Product by Industry* 1961–1971, Ottawa, Statistics Canada, July 1977.

Long Swings in the Dollar: Are They in the Data and Do Markets Know It?

By CHARLES ENGEL AND JAMES D. HAMILTON*

The value of the dollar appears to move in one direction for long periods of time. We develop a new statistical model of exchange rate dynamics as a sequence of stochastic, segmented time trends. We reject the null hypothesis that exchange rates follow a random walk in favor of our model of long swings. Our model also generates better forecasts than a random walk. The specification is a natural framework for assessing the importance of the "peso problem" for the dollar. We nonetheless reject uncovered interest parity. (JEL 431)

Why did the dollar rise so dramatically in the early 1980s only to fall precipitously afterward? Explanations have focused on such factors as the effects of U.S. monetary and fiscal policy on real interest rates (Jeffrey Frankel, 1988, and Martin Feldstein, 1986), lower capital taxes (Olivier Blanchard and Lawrence Summers, 1984), or a "safe haven" effect (Michael Dooley and Peter Isard, 1985).

Important features of the dollar's movements are difficult to reconcile with these explanations under the dominant models of exchange rate determination. Figure 1 plots the number of U.S. dollars required to obtain a German mark, French franc, or British pound over the period 1973:III–1988:I.[1] One is tempted to share Feldstein's (1988, p. 21) summary of these data: "the dollar has experienced three big swings." The first of these is marked by a sustained rise of foreign currencies against the dollar; between the beginning of 1977 and the end of 1979, the mark gained 33 percent against the dollar, the franc gained 21 percent, and the pound 26 percent. This was followed by a five-year surge in the dollar, at the end of which these three European currencies fell 60–90 percent (logarithmically) against the dollar. Early in 1985, foreign currencies once more began to rise, gaining 50–70 percent against the dollar by the end of 1987.

The apparent long swings in the exchange rate pose important challenges for existing theory. In Rudiger Dornbusch's (1976) model, a monetary or fiscal policy change that drives up real interest rates results in a one-time upward jump in the value of the dollar. The dollar is then supposed to depreciate steadily, so as to equate expected returns across countries. Yet as Dornbusch himself noted in 1983,

> The [overshooting] model for the real interest rate does well in explaining that a rise in U.S. interest rates should lead to an appreciation of the real exchange rate. But it fails when it predicts that the real exchange rate should also be depreciating. That has not in fact been happening, and a theory is needed that will explain why the dollar —real or nominal—is both high *and* stuck. [p. 83]

Indeed, the picture seems to have been even worse than Dornbusch painted it—the dollar was high *and rising* for three years prior and two years subsequent to Dornbusch's remarks. Accounting for the gradual, sustained fall in the dollar beginning in 1985 in a way that is consistent with

*Department of Economics, University of Virginia, Charlottesville, VA 22901. This material is based upon work supported by the National Science Foundation under grant no. SES-8720731. We are grateful to Karen Lewis, Anthony Rodrigues, and anonymous referees for helpful suggestions, and to the Federal Reserve Bank of Richmond for assistance with obtaining the data for this project.

[1] The data are normalized so that 1973:III = 1.0 for all three currencies.

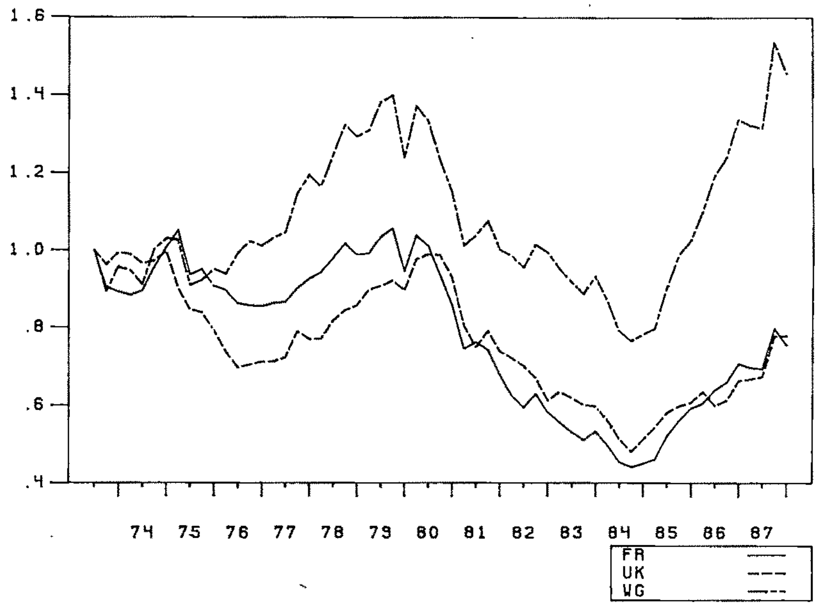


FIGURE 1

the explanation given for its rise is equally problematic.[2]

A further aspect of the apparent long swings that causes difficulty for theories of the exchange rate is that during the period of the strengthening dollar, forward exchange rates (in dollars per unit of foreign currency) were consistently above the spot exchange rate. If the forward rate reflects expectations of future spot rates, then the market appeared to believe over a long period of time that a depreciation of the dollar was imminent. Yet the dollar continued its climb upward until the end of 1984. It could be argued that this forward rate behavior represents an example of William Krasker's (1980) "peso problem." If investors perceived a small probability of a large depreciation, then we might see a forward rate in excess of the current spot rate for a sustained period of time.

For these reasons, it seems useful to formalize the concept of long swings in the exchange rates. What does one mean by long swings, and what magnitudes are plausibly associated with the population parameters? Are long swings a systematic part of the process that generated the data in Figure 1, or a pattern imposed by the eye on the directionless drift of a random walk? If long swings are an accurate description of population dynamics, what sorts of expectations on the part of foreign exchange speculators are consistent with this process? Addressing these questions can provide us with a systematic basis for discussing the issues raised above.

The model we investigate is a special case of that introduced in James Hamilton (1989a).[3] The basic idea is to decompose a

[2] Changes in nominal price differentials between countries were small over this period compared to changes in nominal exchange rates. Thus the real and nominal exchange rates exhibit essentially the same patterns (see Michael Mussa, 1986, on this point). This fact poses additional challenges for theory. In our empirical analysis we focus on the dynamics of nominal exchange rates rather than real exchange rates. When we extend the process to a bivariate system including the nominal interest differential, this permits us to obtain a clean parameterization for testing uncovered interest parity without having to commit ourselves to a model of price level expectation.

[3] Graciela Kaminsky (1988) has also fit Hamilton's model to exchange rate data. She uses monthly data on the pound, whereas we investigate quarterly data on

nonstationary time-series into a sequence of stochastic, segmented time trends. Specifically, we model any given quarter's change in the exchange rate as deriving from one of two regimes, which could correspond to episodes of a rising or falling exchange rate, respectively. The regime at any given date is presumed to be the outcome of a Markov chain whose realizations are unobserved by the econometrician. The task facing the econometrician is to characterize the two regimes and the law that governs the transition between them. These parameter estimates can then be used to infer which regime the process was in at any historical date and provide forecasts for future values of the series.

Our maximum likelihood estimates correspond closely to the visual impressions of Figure 1. In regime 1 the mark is rising 4 percent per quarter against the dollar, the franc 3.3 percent, and the pound 2.6 percent. Regime 2 is associated with quarterly declines in the foreign currencies of -1.2 percent, -2.7 percent, and -3.8 percent, respectively. A given regime is likely to persist for several years, and the econometrically imputed historical change points are close to those the eye is tempted to draw directly from Figure 1.

We perform both Wald tests and likelihood ratio tests that compare the null hypothesis that exchange rates follow a martingale with the segmented trends alternative. In every test but one (the likelihood ratio tests for German data), we reject the martingale hypothesis. The segmented trends model reduces the within sample mean forecast error by 9–14 percent at horizons from two quarters to a year for all three currencies, relative to a random walk specification. Comparable improvements characterize the post-sample forecasts at horizons from one to two quarters. We conclude that long swings in the exchange rate may well be a real feature of the data-generating process.

In exploring the second question posed by our paper—whether markets perceive these swings—we investigate the hypothesis of uncovered interest parity. This hypothesis holds that the nominal interest differential between two countries forecasts future exchange rate changes. This is essentially equivalent to the claim that the three-month forward exchange rate is a rational forecast of the future spot exchange rate. We find no evidence to support this hypothesis in the data. We conclude that either (a) investors did not know the population parameters of the long swings model that generated the historical data, as our rational-expectations calculations assume, or (b) uncovered interest parity does not hold. Big differences in the volatility of exchange rates between the two regimes make it possible that (b) is due to risk aversion on the part of foreign exchange speculators.

The plan of the paper is as follows. Section I sets out the basic model we use to formalize the long swings hypothesis. Section II characterizes our estimation procedure. Empirical results are presented in Section III, while Section IV analyzes the hypothesis of uncovered interest parity. Conclusions are offered in Section V.

I. A Model of Stochastic Segmented Trends

Our model postulates the existence of an unobserved variable (denoted s_t) that takes on the value one or two. This variable characterizes the "state" or "regime" that the process was in at date t. When $s_t = 1$, the observed change in the exchange rate y_t is presumed to have been drawn from a $N(\mu_1, \sigma_1^2)$ distribution, whereas when $s_t = 2$, y_t is distributed $N(\mu_2, \sigma_2^2)$; thus when $s_t = 1$, the trend in the exchange rate is μ_1, whereas when $s_t = 2$, the trend is μ_2.

We further postulate a Markov chain for the evolution of the unobserved state vari-

the pound, mark, and franc. She does not explore the hypothesis of uncovered interest parity and does not devote as much detail to analysis of the random walk hypothesis. She extends Hamilton's model to a case where information besides realized exchange rates is available on the state of the process. Stephan Schulmeister (1987) has also documented the "long swings" phenomenon on the basis of nonparametric statistical analysis.

able:

$$
\begin{aligned}
(1)\qquad & p(s_t=1|s_{t-1}=1)=p_{11}\\
& p(s_t=2|s_{t-1}=1)=1-p_{11}\\
& p(s_t=1|s_{t-1}=2)=1-p_{22}\\
& p(s_t=2|s_{t-1}=2)=p_{22}.
\end{aligned}
$$

The process for s_t is presumed to depend on past realizations of y and s only through s_{t-1}.

Note the variety of behavior that the model allows; in particular, we do not impose that exchange rates are described by long swings. For example, there can be asymmetry in the persistence of the two regimes—upward moves could be short but sharp (μ_1 large and positive, p_{11} small), whereas downward moves could be gradual and drawn out (μ_2 negative and small in absolute value, p_{22} large). Alternatively, the exchange rate change this period could be completely independent of the state that prevailed last period, as in a random walk, if $p_{11}=1-p_{22}$. A third possibility is the long swings hypothesis, which we represent as the claim that μ_1 and μ_2 are opposite in sign and that values for both p_{11} and p_{22} are large.

Our model resembles a standard probability distribution that is called a "mixture of normal distributions." This distribution is a superposition of two (or more) simple normal distributions. A histogram of data drawn from such a distribution would represent the sum of two overlapping bell-shaped curves. The parameters of the distribution would be the mean and variance of each of the simple normal distributions, and a weight given to the first distribution to represent the fraction of realizations that were likely to have been drawn from it. One could use these parameters to calculate the probability that any given observation came from the first distribution. The difference between our model and this mixture of normals is that the draws of y_t in our model are not independent. When we infer the odds that a particular y_t comes from the first distribution, that probability depends on the realizations of y at other times.

II. Maximum Likelihood Estimation of Parameters

The probability law for the data $\{y_t\}$ is summarized by six population parameters,

$$\boldsymbol{\theta}=(\mu_1,\mu_2,\sigma_1,\sigma_2,p_{11},p_{22})'.$$

These parameters are sufficient to describe (a) the distribution of y_t given s_t, (b) the distribution of s_t given s_{t-1} in equations (1), and (c) the unconditional distribution of the state of the first observation:

$$(2)\quad p(s_1=1;\boldsymbol{\theta})\equiv\rho=\frac{(1-p_{22})}{(1-p_{11})+(1-p_{22})};$$

of course $p(s_1=2;\boldsymbol{\theta})=1-\rho$. The joint probability distribution of the observed data for a sample of size T $(y_1,\dots,y_T)$ along with the unobserved states $(s_1,\dots,s_T)$ is then

$$
\begin{aligned}
(3)\quad & p(y_1,\dots,y_T,s_1,\dots,s_T;\boldsymbol{\theta})\\
&=p(y_T|s_T;\boldsymbol{\theta})\cdot p(s_T|s_{T-1};\boldsymbol{\theta})\\
&\cdot p(y_{T-1}|s_{T-1};\boldsymbol{\theta})\cdot p(s_{T-1}|s_{T-2};\boldsymbol{\theta})\cdot\dots\cdot\\
&p(s_2|s_1;\boldsymbol{\theta})\cdot p(y_1|s_1;\boldsymbol{\theta})\cdot p(s_1;\boldsymbol{\theta}).
\end{aligned}
$$

The sample likelihood function could be thought of as the summation of (3) over all possible values of $(s_1,\dots,s_T)$:

$$
\begin{aligned}
(4)\quad & p(y_1,\dots,y_T;\boldsymbol{\theta})\\
&=\sum_{s_1=1}^{2}\cdots\sum_{s_T=1}^{2}p(y_1,\dots,y_T,s_1,\dots,s_T;\boldsymbol{\theta}).
\end{aligned}
$$

In practice we use Hamilton's (1989a) simpler algorithm for evaluation of (4) that does not require 2^T summations.

Given knowledge of the population parameters $\boldsymbol{\theta}$, it is straightforward to characterize the probability that the process was in some particular regime s_t at date t on the basis of information available at the time

$$p(s_t|y_1,\dots,y_t;\boldsymbol{\theta}).$$

We refer to this as the "filter" inference about the probable regime at date t. Alternatively, one can use the full sample of *ex post* available information $(y_1,\dots,y_T)$ to draw an inference about the historical state the process was in at some date t:

$$p(s_t|y_1,\dots,y_T;\boldsymbol{\theta}),$$

which we refer to as the "smoothed" inference about the regime at date t.

Note that unlike the model of a mixture of normal distributions in which the y_t are independent, these probabilities depend at each time on y's that occur at other times. For example, if $s_{t-1}=1$ and p_{11} is high, then y_t is more likely to have been generated from distribution 1; on the other hand, if $s_{t-1}=2$ and p_{22} is high, then y_t is more likely to have been drawn from distribution 2.

Hamilton (forthcoming) showed that first-order conditions for maximization of (4) with respect to $\boldsymbol{\theta}$ characterize the MLE $\hat{\boldsymbol{\theta}}$ as satisfying

$$(5)\quad \hat{\mu}_j=\frac{\sum_{t=1}^{T} y_t\cdot p(s_t=j|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})}{\sum_{t=1}^{T} p(s_t=j|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})}\qquad j=1,2$$

$$(6)\quad \hat{\sigma}_j^2=\frac{\sum_{t=1}^{T} (y_t-\hat{\mu}_j)^2\cdot p(s_t=j|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})}{\sum_{t=1}^{T} p(s_t=j|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})}\qquad j=1,2$$

$$(7)\quad \hat{p}_{11}=\left\{\sum_{t=2}^{T} p(s_t=1,s_{t-1}=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})\right\}\div\left\{\sum_{t=2}^{T} p(s_{t-1}=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})+\hat{\rho}-p(s_1=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})\right\}$$

$$(8)\quad \hat{p}_{22}=\left\{\sum_{t=2}^{T} p(s_t=2,s_{t-1}=2|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})\right\}\div\left\{\sum_{t=2}^{T} p(s_{t-1}=2|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})-\hat{\rho}+p(s_1=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})\right\}.$$

Consider first the intuition behind equation (5). Suppose that the econometrician knows with certainty which observations came from regime 1 and which from regime 2. Then $p(s_t=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})$ is either one or zero for all observations, and the estimated mean for regime 1, $\hat{\mu}_1$, is simply the average value of y_t for those observations known to come from regime 1 (that is, the average of y_t for those dates t for which $p(s_t=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})=1$). If the designation of regimes is not known with certainty, then an observation t whose smoothed probability of coming from regime 1 is 0.3 ($p(s_t=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})=0.3$) is given a weight of 0.3 in constructing the estimate of μ_1 and of 0.7 in constructing the estimate of μ_2.

Similarly, the variance imputed to regime 1 ($\hat{\sigma}_1^2$ in equation 6) is a weighted sum of squared deviations of the observations around the imputed population mean ($\hat{\mu}_1$), with weights again proportional to the probability that any date t's datum was indeed generated from regime 1.

Finally, the estimate of the Markov transition probability (7) is again best understood by first considering the case where designation of regimes is known with certainty ($p(s_t=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})=1$ or 0). Then the estimated Markov transition probability ($\hat{p}_{11}$) is essentially the number of times the transition was made from 1 to 1 as a fraction of the number of times the process had been in state 1 the previous period. In addition, there is a slight adjustment in the denominator for initial conditions. If the process seems to have been in state 1 at date 1 ($p(s_1=1|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})$ large), and yet an ergodic draw from the Markov process (1) is unlikely to be in state 1 ($\hat{\rho}$ in equation

2 is small), then the adjustment favors choosing a larger value for $\hat{p}_{11}$.

We found solutions to equations (5)–(8) by using the EM algorithm developed in Hamilton (forthcoming).

A well-known problem (for example, B. S. Everitt and D. J. Hand, 1981) with estimating parameters for i.i.d. mixtures of normal distributions is that a singularity in the likelihood function arises when, for example, the mean of regime 1 is imputed to equal the value of the realization of the first observation in the sample ($\mu_1 = y_1$) and the variance of regime 1 is permitted to vanish ($\sigma_1 \to 0$). At such a singularity, the likelihood function (4) blows up to infinity. This paper follows Hamilton (1988b) in incorporating a Bayesian prior for the parameters of the two regimes, replacing (5) and (6) with

$$(9)\quad \hat{\mu}_j = \frac{\sum_{t=1}^{T} y_t \cdot p(s_t = j | y_1,\dots,y_T;\hat{\boldsymbol{\theta}})}{\nu + \sum_{t=1}^{T} p(s_t = j | y_1,\dots,y_T;\hat{\boldsymbol{\theta}})}$$

$$(10)\quad \hat{\sigma}_j^2 = \left[\frac{1}{\alpha + (1/2)\cdot \sum_{t=1}^{T} p(s_t = j|y_1,\dots,y_T;\hat{\boldsymbol{\theta}})}\right] \times\left[\beta + (1/2)\cdot \sum_{t=1}^{T} (y_t - \hat{\mu}_j)^2 \cdot p(s_t = j|y_1,\dots,y_T;\hat{\boldsymbol{\theta}}) + (1/2)\cdot \nu \cdot (\hat{\mu}_j)^2\right].$$

This Bayesian approach reproduces the MLE as a special case of the diffuse prior $\nu = \alpha = \beta = 0$. In general, (9) shrinks $\hat{\mu}_j$ toward zero for $j = 1,2$, as if one had, in addition to the observed data $(y_1,\dots,y_T)$, ν additional observations from each regime that took on the value zero. Equation (10) adjusts $\hat{\sigma}_j^2$ toward (β/α), as though one had 2α observations relevant toward this adjustment. Equations (7) and (8) are left as is. The prior thus shifts the MLE estimates in the direction of concluding that there is no difference between the two regimes.

A numerically equivalent way to think about this prior is that one is seeking to maximize not the likelihood function (4) but rather the generalized objective function

$$(11)\quad z(\boldsymbol{\theta}) = \log p(y_1,\dots,y_T;\boldsymbol{\theta}) - \left[(\nu\cdot\mu_1^2)/(2\sigma_1^2)\right] - \left[(\nu\cdot\mu_2^2)/(2\sigma_2^2)\right] - \alpha\log\sigma_1^2 - \alpha\log\sigma_2^2 - \beta/\sigma_1^2 - \beta/\sigma_2^2.$$

Unlike the likelihood function (4), the singularities described above are not a feature of the objective function (11) for α, β, and $\nu > 0$. Monte Carlo simulations reported in Hamilton (1988b) suggest that very modest priors can consistently improve mean squared errors.

III. Empirical Results

A. Maximum Likelihood Estimates

The raw data for this project are an arithmetic average of the bid and asked prices for the exchange rate (in dollars per unit of foreign currency) for the last day of the quarter, beginning with the third quarter of 1973 and ending with the first quarter of 1988. The data are expressed in units of percentage change, denoted y_t^{WG}, y_t^{FR}, and y_t^{UK}.[4]

We estimated the parameter vector $\boldsymbol{\theta}$ for each currency in isolation from the others. Table 1 reports maximum likelihood estimates; the Appendix provides further details.

[4] All series were taken from the data banks compiled by Data Resources, Inc., as of June 1988. The raw data have the DRI series names WGCOOA, WGCOOB, FRCOOA, FRCOOB, UKCOOA, and UKCOOB. Natural logarithms were taken. The data were then first-differenced and multiplied by 100 to express in units of percentage change. The resulting quarterly series (y_t^{WG}, y_t^{FR}, and y_t^{UK}) run from 1973:IV to 1988:I.

TABLE 1—ESTIMATES FIT TO INDIVIDUAL COUNTRY DATA, $y_t = e_t - e_{t-1}$, $t = 73{:}\text{IV}$ TO $88{:}\text{I}$, $e_t = 100$ TIMES THE LOG OF THE EXCHANGE RATE (IN DOLLARS PER UNIT OF FOREIGN CURRENCY)

Parameter	Germany	France	U.K.
μ_1	3.987	3.256	2.627
	(1.230)	(0.967)	(0.872)
μ_2	−1.183	−2.712	−3.752
	(1.480)	(1.367)	(1.139)
p_{11}	0.848	0.822	0.927
	(0.122)	(0.105)	(0.057)
p_{22}	0.928	0.908	0.913
	(0.066)	(0.063)	(0.073)
σ_1^2	17.652	9.991	16.918
	(9.351)	(5.001)	(4.660)
σ_2^2	42.166	36.921	20.247
	(11.242)	(10.252)	(5.841)
$\hat{\rho}$	0.322	0.342	0.542
$p(s_1 = 1 \mid y_1, \ldots, y_T; \hat{\theta})$	0.004	0.000	0.373

Note: Standard errors are in parentheses.

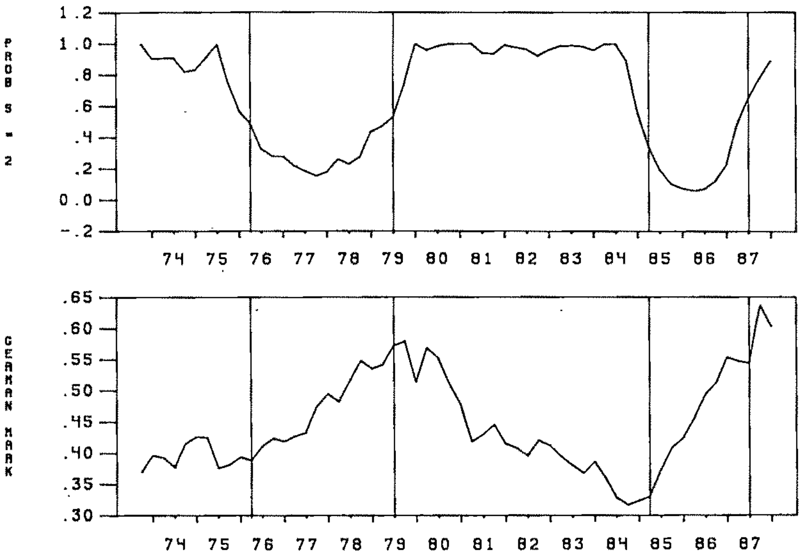


FIGURE 2

The maximum likelihood estimates associate state 1 with a 4 percent quarterly rise in the German mark, a 3.3 percent rise in the franc, and a 2.6 percent rise in the pound. In state 2 the currencies fall by −1.2 percent, −2.7 percent, and −3.8 percent, respectively, with considerably more variability in the exchange rate apparent in state 2 than in state 1.

The bottom panel of Figure 2 plots the exchange rate for the German mark (in $/mark). The top panel plots the smoothed probability that the process was in regime 2 at each date in the sample; that is, $p(s_t = 2 \mid y_1^{WG}, \ldots, y_T^{WG}; \hat{\theta}^{WG})$ is plotted as a function of t. This inference uses the full sample of observations for Germany $(y_1^{WG}, \ldots, y_T^{WG})$ and the maximum likelihood estimates of

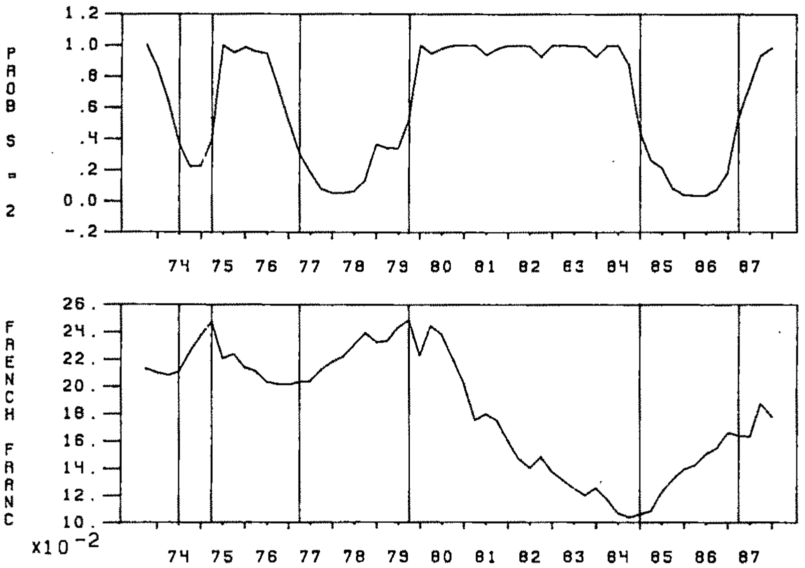


FIGURE 3

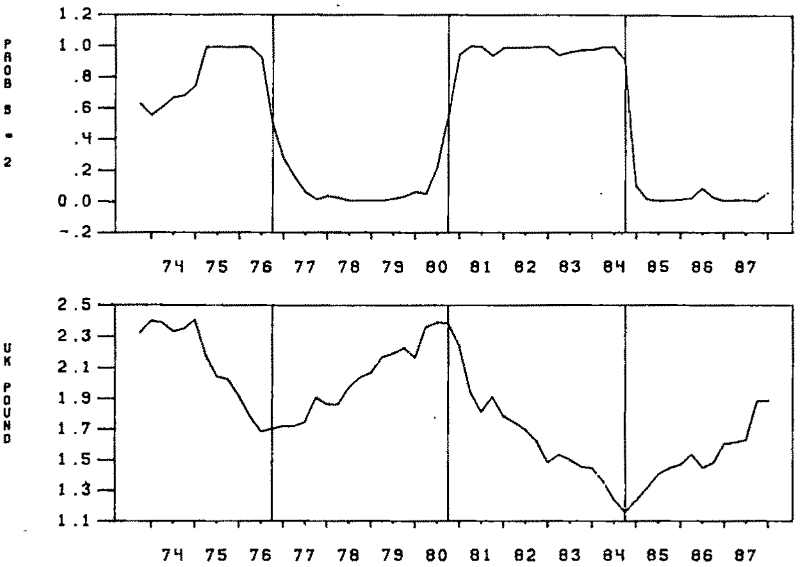


FIGURE 4

parameters $\hat{\theta}^{WG}$ to draw an inference about the state of the process at each date t. The dates at which the econometrician would conclude that the process had switched between regimes (based on $p(s_t = 2|y_1^{WG},\dots,y_T^{WG};\hat{\theta}^{WG}) \gtrless 0.5$) are shown as vertical bars. Similar diagrams for France and the U.K. appear as Figures 3 and 4.

The estimates in Table 1 show that movements in the exchange rate are characterized by long swings. The point estimates of p_{11} range from 0.822 to 0.927, while the estimates of p_{22} go from 0.908 to 0.928. These probabilities indicate that if the system is in either state 1 or state 2, it is likely to stay in that state. Inspection of Figures 2–4 shows that by our estimates the switches between states are infrequent. All of the currencies were in a state of appreciation of the dollar (that is, they were in state 2) from 1980–1984, and were in a state of depreciation of the dollar (state 1) from the end of 1984 to 1987. Thus, our model of long swings tends to match closely what one might be led to believe from casual inspection of Figure 1.[5]

States 1 and 2 are differentiated not only by their means but also by the variances of the conditional distributions. The exchange rate seems to be much more variable when the dollar is appreciating. For the mark and franc, our estimates show that the dollar entered the appreciation stage in the middle of 1987. This assessment is based on the unusual volatility in the exchange rate during that year.

It is straightforward conceptually to generalize this approach to vector processes $\mathbf{y}_t$ (see Hamilton, 1988b). Here we posit that $\mathbf{y}_t|s_t \sim N(\boldsymbol{\mu}_{s_t}, \boldsymbol{\Omega}_{s_t})$. Equations (7), (8), and (9) continue to hold with y_t and μ_j interpreted as the corresponding vectors. Equation (10) is replaced by

$$(12)\quad \hat{\boldsymbol{\Omega}}_j = \left(\frac{1}{\alpha + (1/2)\cdot \sum_{t=1}^{T} p(s_t = j|\mathbf{y}_1,\ldots,\mathbf{y}_T;\hat{\boldsymbol{\theta}})}\right) \times[\boldsymbol{\Lambda} + (1/2)\cdot \sum_{t=1}^{T} (\mathbf{y}_t - \hat{\boldsymbol{\mu}}_j)(\mathbf{y}_t - \hat{\boldsymbol{\mu}}_j)' \cdot p(s_t = j|\mathbf{y}_1,\ldots,\mathbf{y}_T;\hat{\boldsymbol{\theta}}) + (1/2)\cdot \nu \cdot \hat{\boldsymbol{\mu}}_j\hat{\boldsymbol{\mu}}_j']$$

where $(\alpha, \boldsymbol{\Lambda})$ is a multivariate generalization of (α, β) based on the Wishart distribution. Unfortunately, we had little success in using these equations to fit all three currencies to a process driven by a single scalar state variable s_t. The estimates did not correspond well with the individual inferences of any of the three currencies. The behavior of individual exchange rates is determined, of course, not only by events in the United States but also by events in each of the corresponding countries. It appears that treating these three exchange rates as a group is inappropriate because country-specific developments played an important role in the evolution of exchange rates in the 1970s. For this reason, we proceed in our analysis of each of the three countries in insolation from the others.

B. *Testing the Null Hypothesis That Exchange Rates Follow a Random Walk*

An alternative to the segmented trends model is the simple random walk. Michael Mussa (1979), Richard Meese and Kenneth Singleton (1982), Meese and Kenneth Rogoff (1983a, 1983b), and Francis Diebold and James Nason (forthcoming) have all produced evidence that the log of the exchange rate follows a random walk. David Hsieh (1989), however, found evidence consistent with both the earlier random walk conclusions and the predictions of our model, asserting that while there was little or no linear serial dependence in the log of the change in daily exchange rates, there seems to be general nonlinear serial dependence.

There are some knotty methodological problems in testing the null hypothesis that exchange rates follow a random walk against the segmented trends alternative. If one views the null hypothesis as the claim that $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, then under the null hypothesis, the parameters p_{11} and p_{22} are unidentified. Moreover, at the constrained MLE

$$\left(\hat{\mu}_1 = \hat{\mu}_2 = \bar{y} = \sum_{t=1}^{T} y_t/T,\quad \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \sum_{t=1}^{T} (y_t - \bar{y})^2/T\right),$$

[5] Kaminsky (1988) fit Hamilton's model to monthly data on the pound. She assumed constant variances and arrived at parameter estimates for the means and transition probabilities, as well as inference about historical switch points, that are comparable to those we find for quarterly data.

TABLE 2—TESTS OF THE NULL HYPOTHESIS THAT EXCHANGE RATES FOLLOW A MARTINGALE AGAINST THE ALTERNATIVE OF SEGMENTED TRENDS

	H_0': $p_{11}=1-p_{22}$		H_0'': $\mu_1=\mu_2$	
Country	Wald Test	Likelihood Ratio Test	Wald Test	Likelihood Ratio Test
Germany	24.70	2.49	8.64	3.00
	(0.00)	(0.11)	(0.00)	(0.08)
France	29.45	5.70	12.65	6.10
	(0.00)	(0.02)	(0.00)	(0.01)
U.K.	61.26	6.15	25.80	8.85
	(0.00)	(0.01)	(0.00)	(0.00)

Note: All statistics are asymptotically $\chi^2(1)$. Asymptotic p-values are in parentheses.

the derivative of the likelihood function with respect to μ_1 or σ_1 is identically zero. These difficulties combine features of the problems discussed by Mark Watson and Robert Engle (1985) and Lung-Fei Lee and Andrew Chesher (1986). The information matrix is singular under the null, and the standard regularity conditions for establishing asymptotically valid tests of H_0 do not hold in this case.

In this paper we sidestep these issues by focusing on the following slightly more general null hypothesis:

$$H_0': \quad p_{11}=1-p_{22}$$
$$\mu_1\neq\mu_2$$
$$\sigma_1\neq\sigma_2.$$

Note that under H_0', the distribution of s_t is independent of s_{t-1}; [from (1), the probability that $s_t=1$ is p_{11} regardless of whether $s_{t-1}=1$ or 2]. Changes in the exchange rate under H_0' thus comprise an i.i.d. sequence with individual densities given by the following mixture of two normals:

$$p(y_t;\boldsymbol{\theta})=\frac{p_{11}}{\sqrt{2\pi}\,\sigma_1}\exp\left[\frac{-(y_t-\mu_1)^2}{2\sigma_1^2}\right]$$
$$+\frac{(1-p_{11})}{\sqrt{2\pi}\,\sigma_2}\exp\left[\frac{-(y_t-\mu_2)^2}{2\sigma_2^2}\right].$$

We can then hope to test H_0' against the alternative that $p_{11}\neq 1-p_{22}$, using standard distribution theory, since under H_0' the parameters μ_1, μ_2, σ_1, σ_2, p_{11}, and p_{22} are all identified.

We report two tests of H_0', the fist being a Wald test. Let $\hat{\text{var}}(\hat{p}_{jj})$ denote the asymptotic variance of $\hat{p}_{jj}$ (as estimated from the inverse of the negative of the matrix of second derivatives of (11)) and $\hat{\text{cov}}(\hat{p}_{11},\hat{p}_{22})$ the asymptotic covariance. Then under H_0',

$$(13)\qquad \frac{[\hat{p}_{11}-(1-\hat{p}_{22})]^2}{[\hat{\text{var}}(\hat{p}_{11})+\hat{\text{var}}(\hat{p}_{22})+2\hat{\text{cov}}(\hat{p}_{11},\hat{p}_{22})]}\approx\chi^2(1).$$

Column 1 of Table 2 reports this Wald test statistic for the three currencies. The 5 percent critical value for a $\chi^2(1)$ variate is 3.84, implying overwhelming rejection of H_0' for all three currencies.

We also tested H_0' by using a likelihood ratio test.[6] This statistic compares the value of the objective function achieved by the estimates in Table 1 with the largest value achievable when estimated subject to the constraint $p_{11}=(1-p_{22})$. The latter estimation is a straightforward application of estimating parameters for an i.i.d. mixture of two normals; we used the EM algorithm described in Everitt and Hand (1981, pp.

[6] A. Ronald Gallant (1987, p. 219) argues that the likelihood ratio test is apt to be more robust than Wald tests in a nonlinear model such as this one.

36–37) for this purpose, with Bayesian correction as in (9) and (10). Twice the difference in the objective function (11) between the constrained and unconstrained estimates is reported in column 2 of Table 2, and is presumed asymptotically to have a $\chi^2(1)$ distribution. The magnitude of the difference between the Wald test statistics and the likelihood ratio test statistics is disconcerting.[7] Still, at least in the case of France and the U.K., the rejection of H_0' continues to be fairly decisive.

It is also interesting to test the hypothesis H_0'': $\mu_1 = \mu_2$. Under this hypothesis, the exchange rate follows a stochastic process as described in Section I, with the mean rate of depreciation the same in both states. If $\mu_1 = \mu_2$, but $\sigma_1 \neq \sigma_2$, the states have the same mean rate of depreciation but different variances. A Wald statistic for testing H_0'' is given by

$$\frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{\hat{\text{var}}(\hat{\mu}_1) + \hat{\text{var}}(\hat{\mu}_2) - 2\hat{\text{cov}}(\hat{\mu}_1, \hat{\mu}_2)} \approx \chi^2(1).$$

The statistics are reported in column 3 of Table 2. Column 4 reports an analogous likelihood ratio test. The means of the two states are different.

We thus conclude that movements in the dollar are described by long swings. The dollar enters stages in which it appreciates or depreciates and it remains in such stages for years. The expected length of state i is $1/(1 - p_{ii})$. State 1 is expected to persist for seven quarters for Germany, six for France, and fourteen for the U.K. On average state 2 lasts fourteen quarters for the mark, eleven for the franc, and twelve for the pound.

C. *Forecasting*

As further evidence on the random walk hypothesis, we calculated the in-sample and post-sample forecast errors for the segmented trends model in comparison to those of the random walk specification.

Consider first in-sample forecasts. If one takes the MLE $\hat{\boldsymbol{\theta}}$ (about which the full sample $y_1, \ldots, y_T$ was used to draw inference) as known at date t, then the forecast one would make on the basis of observation of y through date t and on the basis of knowledge of $\hat{\boldsymbol{\theta}}$ is given by[8]

$$(14)\quad E\left[y_{t+j}|y_t, y_{t-1}, \ldots, y_1; \hat{\boldsymbol{\theta}}\right] = \hat{\mu}_2 + \left\{\hat{\rho} + (-1 + \hat{p}_{11} + \hat{p}_{22})^j \cdot \left[p(s_t = 1|y_1, \ldots, y_t; \hat{\boldsymbol{\theta}}) - \hat{\rho}\right]\right\} \cdot \{\hat{\mu}_1 - \hat{\mu}_2\},$$

with $\hat{\rho}$ given in (2). Letting $\hat{y}_{t+j|t}$ denote the expression in (14), we calculated k-period ahead forecasts of the level of the log of the exchange rate

$$\hat{e}_{t+k|t} = e_t + \hat{y}_{t+1|t} + \hat{y}_{t+2|t} + \cdots + \hat{y}_{t+k|t}$$

and calculated the average squared value of the forecast error,

$$\sum_{t=1}^{T-k} (\hat{e}_{t+k|t} - e_{t+k})^2/(T-k),$$

for forecast horizons (k) of one through four quarters.

The top panel of Table 3 compares these forecast errors with those of a random walk specification, whose forecasts are given by $\hat{e}_{t+k|t} = e_t + k \cdot \bar{y}$, where $\bar{y} = \Sigma_{t=1}^{T} y_t / T$. Note that the variance of the latter forecast error should, if the random walk specification is correct, rise linearly with the forecast horizon k. The actual MSE's for the random walk specification in Table 3 perform more poorly than this, owing to positive autocorrelation in y_t at lags one through

[7] This seems due in part to asymmetry in the likelihood surface for increases and decreases in p_{jj}. For values of p_{jj} above $\hat{p}_{jj}$, the Wald approximation slightly understates the true curvature of the likelihood function, whereas for $p_{jj} < \hat{p}_{jj}$, the likelihood function quickly becomes much flatter than the Hessian evaluated at $\hat{p}_{jj}$ predicts.

[8] See equation (3.2) in Hamilton (1989a).

TABLE 3—IN-SAMPLE AND POST-SAMPLE MEAN SQUARED FORECAST ERROR AT HORIZONS FROM ONE TO FOUR QUARTERS OF SEGMENTED TRENDS MODEL AND RANDOM WALK WITH DRIFT

A. In-Sample Mean Squared Forecast Errors

Country	Forecast Horizon (Quarters)			
(Model)	1	2	3	4
Germany				
(Random Walk)	38.01	83.79	130.84	199.18
(Segmented Trend)	36.48	76.39	113.55	174.93
(Percent Improvement)	4 Percent	9 Percent	13 Percent	12 Percent
France				
(Random Walk)	34.37	82.96	143.99	220.88
(Segmented Trend)	31.27	72.13	124.31	194.76
(Percent Improvement)	9 Percent	13 Percent	14 Percent	12 Percent
United Kingdom				
(Random Walk)	29.10	76.06	124.45	187.26
(Segmented Trend)	25.11	65.95	107.82	170.36
(Percent Improvement)	14 Percent	13 Percent	13 Percent	9 Percent

B. Post-Sample Mean Squared Forecast Errors

Country	Forecast Horizon (Quarters)			
(model)	1	2	3	4
Germany				
(Random Walk)	54.58	141.33	251.62	406.49
(Segmented Trend)	50.44	133.37	245.77	409.85
(Percent Improvement)	8 Percent	6 Percent	2 Percent	−1 Percent
France				
(Random Walk)	52.47	145.59	266.34	426.76
(Segmented Trend)	46.80	134.32	255.33	427.01
(Percent Improvement)	11 Percent	8 Percent	4 Percent	−0 Percent
United Kingdom				
(Random Walk)	42.35	117.54	186.68	270.43
(Segmented Trend)	35.11	98.40	161.28	252.61
(Percent Improvement)	17 Percent	16 Percent	14 Percent	7 Percent

Notes: A. In-Sample Forecast Errors. In each case, the population parameters were estimated by using data from $t = 1973\text{:IV}$ to 1988:I and mean squared errors are those associated with forecasts for dates $t = 1973\text{:IV} + k$ to 1988:I where k is the forecast horizon.

B. Post-Sample Forecast Errors. In each case, the population parameters were estimated by using data from $t = 1973\text{:IV}$ to 1983:IV and mean squared errors are those associated with forecasts for dates $t = 1984\text{:I}$ to 1988:I.

three[9] and to the fact that $\bar{y}$, the mean of observations 1 through T, is not quite the same as the mean of observations k though T. The improvement in forecasting at horizons of two to four quarters offered by the segmented trends specification is 9–14 percent for all three currencies.

It is worth noting that our model is doing more than just mimicking an AR(1) specification for exchange rate changes. An AR(1) model has an in-sample one-quarter ahead R^2 of 8 percent for the U.K., 4 percent for France, and less than 1 percent for Germany and offers virtually no improvement in forecasting at longer horizons.

To evaluate the post-sample forecasting performance of the model, we reestimated

[9] Recall John Cochrane's (1988, p. 906) result that

$$\mathrm{Var}(e_{t+k} - e_t) = k\left[\mathrm{Var}(e_{t+1} - e_t) + 2\sum_{j=1}^{k-1}((k-j)/k)\cdot\mathrm{Cov}(y_t, y_{t-j})\right].$$

the parameters with data only up to the end of 1983. We chose the end of 1983 because the major turning point in the dollar that occurred in 1985 had not yet happened. Hence, the entire period of the dollar depreciation of 1985–1987 was not used for estimating parameters. Furthermore, with our out-of-sample forecasts our model must meet the challenge of picking out the turning point.

The parameter estimates for the truncated sample are similar to those of the full sample; using only data through 1983, there is evidence in favor of the long swings hypothesis.

The bottom panel of Table 3 compares the post-sample mean squared error of the forecasts of our model with that of a random walk (with the drift term estimated from data through the end of 1983). The forecasts for the segmented trends model were calculated as in equation (14), but using the parameter estimates from the restricted sample. We found that our model generally outperformed the random walk, particularly at short forecasting horizons.

Table 3 follows Mussa (1979) and Meese and Singleton (1982) in including a drift term in the random walk; by contrast, Meese and Rogoff (1983a, 1983b) and Diebold and Nason (1989) set the drift term a priori to zero. The zero-drift random walk specification has a forecasting performance over 1984:I–1988:I that significantly beats both our specification and the random walk with drift. Some would interpret this as evidence in favor of the random walk hypothesis. We would instead argue that the superiority during 1984–1988 of the random walk without drift over the random walk with drift offers conclusive evidence that exchange rates do *not* follow a random walk, with or without drift! If the data really followed a driftless random walk, then differences in post-sample forecasts between the two random-walk specifications should be entirely due to error in estimating the drift term. Conventional statistical tests lead to clear rejection of the null hypothesis that the exchange rate data come from a random walk with the same drift term before and after 1984 (see also Table 4 below). The driftless random walk is just a special case of this rejected hypothesis. Imposing a particular numerical value for the drift (in this case, zero) is of course going to improve the fit over selected subsamples, but cannot salvage the model as a specification that describes the complete sample. An apparent break in parameter values over particular subsamples may be an important feature that accounts for the results of Meese and Rogoff and Diebold and Nason, and it is precisely the feature of the data that our long swings representation is attempting to model.

We conclude that the accumulated evidence from the Wald tests, likelihood ratio tests, and forecasting performance favors the segmented-trends specification over the random walk.

D. *Specification Testing*

This section presents specification tests that fall into four broad groups. The first group explores the forecastability of the one-quarter-ahead in-sample forecast errors. Second, we consider tests based on the work of Whitney Newey (1985), George Tauchen (1985), and Halbert White (1987) that examine the null hypothesis that the score statistics are serially uncorrelated. Third, we perform Lagrange multiplier tests for various sorts of dynamic misspecification. Finally, we split the sample at the end of 1979 and again at the end of 1983 and perform likelihood ratio tests for changes in the stochastic process governing exchange rates.

Forecasting Tests. Our forecasting tests divide the one-quarter-ahead forecasts of our model (expression 14) by their conditional standard deviation:

$$\hat{\sigma}_{\hat{y}_{t+1|t}} = \Big\{\big\{\big[\hat{\mu}_1^2 + \hat{\sigma}_1^2\big]\cdot\hat{p}_{t+1|t} + \big[\hat{\mu}_2^2 + \hat{\sigma}_2^2\big]\cdot\big(1-\hat{p}_{t+1|t}\big)\big\} - \big\{\hat{\mu}_1\cdot\hat{p}_{t+1|t} + \hat{\mu}_2\cdot\big(1-\hat{p}_{t+1|t}\big)\big\}^2\Big\}^{1/2},$$

where

$$\hat{p}_{t+1|t} = (1-\hat{p}_{22}) + (-1+\hat{p}_{11}+\hat{p}_{22}) \cdot p(s_t = 1|y_1,\dots,y_t;\hat{\theta}).$$

The resulting standardized one-period-ahead forecasts errors ($\hat{u}_{t+1}$) should be unforecastable with any time t variables. We calculated many regressions of these errors on their own lagged values, on lagged values of the log changes in exchange rates, and on various combinations of the squares and cross-products of these variables. In no case did a joint test of a zero intercept and zero slope coefficients reject the null hypothesis. For example, a regression of $\hat{u}_t$ on a constant, $\hat{u}_{t-1}$ and $\hat{u}_{t-1}^2$, for $t = 74{:}\text{II}-88{:}\text{I}$ yields $F(3,53)$ statistics whose p-values are (0.90), (0.84), and (0.70) for the three currencies. The smallest p-value in the two dozen regressions we looked at was (0.42) for the regression of $\hat{u}_t$ on $\hat{u}_{t-j}$ and $\hat{u}_{t-j}^2$ for $j = 1,2,3,4$ for the U.K.

Hsieh (1989) used the test of William Brock, W. Davis Dechert, and José Scheinkman (1987) to search for general nonlinear dependence in a time-series. He found evidence of significant dependence in daily data for several currencies. We repeated these tests on our standardized residuals.[10] We varied N (the dimension of "N-histories") between 2 and 6 (in steps of 1) and ℓ (the distance measure, in standard deviations of the data) between 0.5 and 1.5, in steps of 0.25. We found no evidence of serial dependence in the standardized residuals. However, in contrast to Hsieh's analysis of daily data, there is little evidence from this test of nonlinear dependence in the raw quarterly exchange rate changes.

Score tests. White (1987) noted that if a maximum likelihood model is correctly specified, the score statistics (the derivative of the conditional log likelihood of the t-th observation) should be serially uncorrelated. Hamilton (1989b) showed how White's results may be used to construct tests for possible alternatives to the Markov switching model. Table 4 applies these tests to the random walk specification, and Table 5 to our long swings model.

By considering the score with respect to the mean, a White test for autocorrelation can be constructed (which essentially tests for the correlation of the score at time t with respect to μ_i and the score at time $t-1$ with respect to μ_j, for $i,j = 1,2$.) Table 4 provides evidence of autocorrelation in the raw data for the U.K., while Table 5 finds no evidence of autocorrelation left over after fitting the long swings model.

An ARCH test can be implemented by examining the serial correlation properties of the scores with respect to σ_i^2, $i = 1,2$. Table 5 shows that the test for ARCH for the U.K. is significant at the 5 percent level. However, Hamilton (1989b) concluded from Monte Carlo simulations that "For a sample as small as 50 observations, one might be better off using the 1 percent critical value from the asymptotic distributions (rather than the 5 percent value) as a rough guide for a 5 percent small-sample test based on the Newey-Tauchen-White specification tests or Lagrange multiplier tests for misspecification of the variance." By this standard, the null hypothesis of no ARCH should not be rejected.

The Markov assumption that $p(s_t = i)$ depends only on the state at time $t-1$ can be tested against the alternatives that it depends on the state at earlier times or that it depends on the realizations of the data y_{t-1}. This test checks whether the score with respect to the transition probabilities can be predicted by the corresponding lagged score or the score with respect to the mean. Table 5 shows that the Markov specification cannot be rejected for any currency.

Lagrange Multiplier Tests. Tables 4 and 5 also present Lagrange multiplier tests of the random walk and long swings specifications. We tested against the alternatives that there is omitted autocorrelation only in state 1, autocorrelation only in state 2, and autocorrelation across regimes. These produced the same conclusions as the White tests for autocorrelation.

[10] We calculated these statistics by using computer code kindly distributed to us by W. Davis Dechert.

TABLE 4—TESTS OF NULL HYPOTHESIS THAT PERCENT CHANGES IN EXCHANGE RATES ARE i.i.d. GAUSSIAN

Test	Germany	France	U.K.
White Test for	0.28	2.15	**[6.02]**
Autocorrelation ($\chi^2(1)$)	(0.60)	(0.14)	(0.01)
White Test for	0.32	0.02	0.39
ARCH ($\chi^2(1)$)	(0.57)	(0.89)	(0.53)
LM Test for	0.28	2.14	**[6.02]**
Autocorrelation ($\chi^2(1)$)	(0.60)	(0.14)	(0.01)
LM Test for	0.68	0.00	0.32
ARCH ($\chi^2(1)$)	(0.41)	(1.00)	(0.57)
LM Test for Shift in Mean	**[4.14]**	**[6.70]**	**[3.92]**
79:IV–82:IV ($\chi^2(1)$)	(0.04)	(0.01)	(0.05)
LM Test for Shift in Mean	**[4.53]**	**[7.69]**	**[10.19]**
85:II–83:I ($\chi^2(1)$)	(0.03)	(0.01)	(0.00)

Notes: All statistics are asymptotically $\chi^2(1)$ [5 percent critical value = 3.84; 1 percent critical value = 6.63]. Brackets [ ] denote significant at 5 percent level. Asymptotic p-values are in parentheses.

TABLE 5—TESTS OF NULL HYPOTHESIS THAT EXCHANGE RATES FOLLOW THE LONG SWINGS MODEL

Test	Germany	France	U.K.
White Test for	2.56	4.23	4.61
Autocorrelation ($\chi^2(4)$)	(0.63)	(0.38)	(0.33)
White Test for	3.52	7.26	**[10.55]**
ARCH ($\chi^2(4)$)	(0.47)	(0.12)	(0.03)
White Test of Markov	2.73	4.47	1.59
Specification ($\chi^2(4)$)	(0.60)	(0.35)	(0.81)
LM Test for Autocorrelation	0.00	3.26	3.02
in Regime 1 ($\chi^2(1)$)	(1.00)	(0.07)	(0.08)
LM Test for Autocorrelation	1.44	0.24	0.30
in Regime 2 ($\chi^2(1)$)	(0.23)	(0.62)	(0.58)
LM Test for Autocorrelation	0.59	1.06	0.65
Across Regimes ($\chi^2(1)$)	(0.44)	(0.30)	(0.42)
LM Test for	1.27	0.10	**[4.47]**
ARCH ($\chi^2(1)$)	(0.26)	(0.75)	(0.03)
LM Test for Shift in Mean	1.66	1.92	1.50
79:IV–82:IV ($\chi^2(1)$)	(0.20)	(0.17)	(0.22)
LM Test for Shift in Mean	**[11.00]**	**[12.24]**	2.28
85:II–88:I ($\chi^2(1)$)	(0.00)	(0.00)	(0.13)

Notes: The first three statistics are asymptotically $\chi^2(4)$ [5 percent critical value = 9.49; 1 percent critical value = 13.28]. All other statistics are asymptotically $\chi^2(1)$ [5 percent critical value = 3.84; 1 percent critical value = 6.63]. Brackets [ ] denote significant at 5 percent level. Asymptotic p-values are in parentheses.

Tables 4 and 5 also report the results of LM tests for ARCH. For the alternative to the long swings model, the variance at time t, h_t, is modeled as

$$h_t = \gamma_{s_t}\left[1+\frac{\xi(y_{t-1}-\mu_{s_{t-1}})^2}{\gamma_{s_{t-1}}}\right], \qquad i=1,2$$

(see Hamilton, 1989b). Under the null hypothesis of no ARCH, $\xi = 0$, and $\gamma_{s_t} = \sigma_{s_t}^2$. Again, we found some evidence of ARCH for the U.K., but we would probably not consider it significant at the 5 percent level given our number of observations.

We can also use the Lagrange multiplier principle to test whether the mean of the

TABLE 6—LIKELIHOOD RATIO TESTS FOR WHETHER ALL PARAMETERS OF THE LONG SWINGS MODEL CHANGE AT SPECIFIED DATES

Date of Sample Break	Germany	France	U.K.
1979:IV	8.343	12.313	4.846
	(0.21)	(0.06)	(0.56)
1983:IV	4.303	8.050	3.775
	(0.64)	(0.23)	(0.71)

Notes: All variables are asymptotically $\chi^2(6)$. Asymptotic p-values are in parentheses.

process shifted over any subsample. When we applied this test to the random walk specification, we found evidence for all three currencies of a change in the drift associated with the change in U.S. Federal Reserve operating procedures during October 1979–October 1982 (see Table 4). By contrast, allowing for a separate mean for this subperiod does not make a statistically significant contribution to the long swings model (Table 5). Thus one feature of the data that is inconsistent with a random walk that the long swings model captures is the persistent tendency for the dollar to appreciate during the three-year period in which the Fed targeted nonborrowed reserves.

We also tested for a permanent break in the mean of the series for all possible change points in the sample. It is interesting that for all three currencies and for both the random walk and long swings specifications, the largest value of this statistic comes within one quarter of 1985:II. Table 4 reveals evidence of a break in the process after 1985 that is not captured by the random walk. Table 5 suggests that the long swings model is able to capture this break in the case of the U.K. but not in the case of Germany and France.

Likelihood Ratio Tests. We also tested for shifts in the stochastic process by performing likelihood ratio tests for joint changes in all the parameters at the end of 1979 and at the end of 1983. Table 6 shows that we cannot reject the null of no shift at the 5 percent level for any currency at either date.

IV. Testing the Hypothesis of Uncovered Interest Parity

We now turn to the second question posed by our paper—Is this apparent forecastability of exchange rates reflected in intercountry interest differentials? Uncovered interest parity posits that a three-month Eurodollar account should yield the same return expected by converting the dollars to marks, holding these marks in a Euromark account for three months, and converting back into dollars at the then-prevailing exchange rate:

$$(15)\quad i_t^{US} = i_t^{WG} + E_t\left(e_{t+1}^{WG} - e_t^{WG}\right) + u_t.$$

Here i_t is the return on a Eurocurrency account for the specified currency, e_t is the log of the exchange rate (in dollars per unit of foreign currency), and u_t is a disturbance term that reflects measurement and specification error.

Let the log of the forward exchange rate be f_t^{WG} dollars per mark. A pure arbitrage opportunity exists unless

$$(16)\qquad i_t^{US} = i_t^{WG} + f_t^{WG} - e_t^{WG}.$$

Thus, the hypothesis of uncovered interest parity (15) is essentially equivalent to the hypothesis that the forward rate is an unbiased predictor of the future spot rate

$$f_t^{WG} = E_t e_{t+1}^{WG} + u_t.$$

We report our results in terms of testing uncovered interest parity rather than of testing the risk neutrality of the forward currency market, though the two tests are conceptually the same.

Suppose that investors know the population parameter $\mathbf{\theta}$ of the segmented trends model and further observe the value of s_t, which governed the mean of the exchange rate change between $t-1$ and t. When $s_t = 1$, then the change in the exchange rate between t and $t+1$ will be drawn from a $N(\mu_1, \sigma_1^2)$ distribution with probability p_{11} and from a $N(\mu_2, \sigma_2^2)$ distribution with

TABLE 7—INTEREST DIFFERENTIALS AS PREDICTED BY (a) THE UNIVARIATE MLEs FOR EACH COUNTRY'S EXCHANGE RATE (TABLE 1 PARAMETERS), AND (b) THE BIVARIATE MLEs FOR EACH COUNTRY'S EXCHANGE RATE TOGETHER WITH THAT COUNTRY'S INTEREST DIFFERENTIAL (TABLE 8 PARAMETERS)

	Germany	France	U.K.
1. Predicted Value for $i_t^{US}-i_t^*$			
When $s_t=1$ Based on			
(a) Univariate Estimate of	3.203*	2.196*	2.159*
$p_{11}\mu_1+(1-p_{11})\mu_2$	(1.150)	(0.934)	(0.904)
(b) Bivariate Estimate of	0.542*	−0.249*	−0.282*
$\mu_1(2)$	(0.061)	(0.118)	(0.090)
2. Predicted Value for $i_t^{US}-i_t^*$			
When $s_t=2$ Based on			
(a) Univariate Estimate of	−0.811	−2.160	−3.199
$p_{22}\mu_2+(1-p_{22})\mu_1$	(1.335)	(1.214)	(1.001)
(b) Bivariate Estimate of	1.171	−1.228	−1.425
$\mu_2(2)$	(0.086)	(0.238)	(0.174)
3. Predicted Value for Change in $i_t^{US}-i_t^*$			
When State Changes from 2 to 1 Based on			
(a) Univariate Estimate of	4.014*	4.356*	5.367*
$(-1+p_{11}+p_{22})(\mu_1-\mu_2)$	(1.597)	(1.447)	(1.182)
(b) Bivariate Estimate of	−0.629*	0.979*	1.143*
$\mu_1(2)-\mu_2(2)$	(0.102)	(0.249)	(0.193)

Notes: Standard errors are in parentheses. The standard error for a nonlinear function $h(\boldsymbol{\theta})$ of the $(p\times 1)$ parameter vector $\boldsymbol{\theta}$ was calculated as the square root of $[h_\theta(\hat{\boldsymbol{\theta}})]'\hat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})[h_\theta(\hat{\boldsymbol{\theta}})]$ where $h_\theta(\hat{\boldsymbol{\theta}})$ denotes the $(p\times 1)$ vector of derivatives of the function $h(\cdot)$ with respect to the elements of $\boldsymbol{\theta}$, evaluated at the MLE $\hat{\boldsymbol{\theta}}$, and $\hat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ denotes the $(p\times p)$ estimated variance-covariance matrix of $\hat{\boldsymbol{\theta}}$.

*Indicates that the 95 percent confidence intervals for (a) and (b) fail to overlap.

$\mu_1(j)$ denotes the jth element of the vector $\boldsymbol{\mu}_1$ in Table 8; thus $\mu_1(2)$ is the mean interest differential when the process is in state 1, and $\mu_2(2)$ is the mean interest differential when the process is in state 2.

probability $(1-p_{11})$. Thus when $s_t=1$, investors would forecast a change in the exchange rate between t and $t+1$ of

$$E_t(e_{t+1}^{WG}-e_t^{WG}) = p_{11}\mu_1+(1-p_{11})\mu_2, \tag{17a}$$

whereas when $s_t=2$, their forecast would be

$$E_t(e_{t+1}^{WG}-e_t^{WG}) = p_{22}\mu_2+(1-p_{22})\mu_1. \tag{17b}$$

Substituting (17) into (15) gives

$$i_t^{US}-i_t^{WG}=p_{11}\mu_1+(1-p_{11})\mu_2+u_t \quad \text{when } s_t=1 \tag{18a}$$

$$=p_{22}\mu_2+(1-p_{22})\mu_1+u_t \quad \text{when } s_t=2. \tag{18b}$$

Rows (1a) and (2a) of Table 7 present the predicted value for the interest differentials based on the univariate maximum likelihood estimates for each country's exchange rate.

The predictions for state 2 (row (2a)) are particularly interesting. State 2 is the state in which the dollar appreciates. During the period of the dollar appreciation of 1980–1984, the forward rate generally exceeded the current spot rate,[11] implying under uncovered interest parity that markets expected a depreciation of the dollar. One way to reconcile this finding with rationality of expectations is to argue that the econometrician faces a "peso problem" (see Robert Cumby and Maurice Obstfeld, 1984;

[11] See for example George Evans (1986), Jeffrey Frankel and Kenneth Froot (1987, 1988), Eduardo Borensztein (1987), and Robert Cumby (1988).

Robert Hodrick and Sanjay Srivastava, 1984; and Hodrick, 1987). When there is a small probability of a large depreciation, the forward rate may consistently predict a depreciation while none occurs.

This possibility is in principle allowed by equation (17b). Even when the process is in state 2, in which the dollar is more likely than not to appreciate ($\mu_2 < 0$, $p_{22} > 0.5$), the expected change in the exchange rate could be positive if the product of $(1 - p_{22})$ and μ_1 is sufficiently large. However, the probability of a depreciation could not have been large, because the appreciation stage lasted so long. The probability of a depreciation given that we are in state 2 is $1 - p_{22}$; our estimate is 0.072 for the mark. The value of μ_1 for Germany, 3.987, is large but not large enough to justify a positive interest differential. Our calculations suggest that a substantial negative differential of $0.928(-1.183) + 0.072(3.987) = -0.811$ was warranted despite the potential "peso" effect. From mid-1980 to mid-1984 the German mark was in state 2 and yet the U.S.-German interest differential was invariably positive. There is thus *prima facie* evidence against the joint hypothesis of uncovered interest parity and rational expectations. We have allowed for a "peso problem," but the evidence indicates that the probability of leaving state 2 once you are in it is so small that a positive interest differential is unwarranted.[12] Notwithstanding, a 95 percent confidence interval for the predicted interest differential does include positive values —the interval ranges from -3.427 to 1.805.

Essentially the same conclusion holds for the U.K. The predicted interest differential in state 2 is -3.20—indeed, the upper end of the 95 percent confidence interval for the predicted interest differential (-5.161 to -1.237) is negative—while the U.S.-U.K. interest differential was almost always positive from the end of 1980 to 1984. This is again a period when our estimates imply that the exchange rate was surely in state 2.

With France, we cannot make such a bold statement. It is still true that when we are in state 2 our univariate estimates of the exchange rate indicate the interest differential should be negative. However, the U.S.-French three-month interest differential was frequently negative during 1980–1984. So there is not a simple, clear-cut case against interest parity in the case of the dollar/franc relationship.

Of course, one could still try to salvage the peso story by postulating a possible depreciation of the dollar that is more dramatic than that associated with regime 1. According to this view, there is perhaps a third possible regime of violent depreciation, which was never observed in the sample. The problem with this view is that, under rational expectations, this massive depreciation has to be regarded as an extremely unlikely event—it did not happen once in 58 observations. Suppose we therefore take the probability of moving into this regime, p_{23}, as less than 0.02. For such a remote event to be able to change the calculations in row (2a) of Table 7 from negative to positive, the *quarterly* depreciation of the dollar in state 3 would have to be 40 percent (logarithmically) against the mark, 108 percent against the franc, and 160 percent agains the pound!

We now explore the hypothesis of uncovered interest parity by examining the joint behavior of exchange rates and interest rates. Expressions (18) predict that the interest differential at date t should have one of two means, selected by the same state variable s_t that governed the realization of the exchange rate change observed at t. Consider then the two-dimensional vector

$$\mathbf{y}_t \equiv \left[\left(e_t^{WG} - e_{t-1}^{WG}\right), \left(i_t^{US} - i_t^{WG}\right)\right]'.$$

The model holds that this vector comes from one of two distributions:

$$\mathbf{y}_t|(s_t = 1) \sim N\left(\begin{bmatrix} \mu_1 \\ p_{11}\mu_1 + (1 - p_{11})\mu_2 \end{bmatrix}, \boldsymbol{\Omega}_1\right)$$

$$\mathbf{y}_t|(s_t = 2) \sim N\left(\begin{bmatrix} \mu_2 \\ p_{22}\mu_2 + (1 - p_{22})\mu_1 \end{bmatrix}, \boldsymbol{\Omega}_2\right),$$

where we put no restrictions on the vari-

[12] This is reminiscent of the argument made by Frankel (1985) that a rational stochastic bubble could not explain the behavior of the dollar.

TABLE 8—ESTIMATES FIT TO $y_t = [(e_t - e_{t-1}),(i_t^{US} - i_t^{WG})]'$, $t = 73\text{:IV}–88\text{:I}$,
e_t = 100 TIMES THE LOG OF THE EXCHANGE RATE
(IN DOLLARS PER UNIT OF FOREIGN CURRENCY),
i_t = INTEREST RATE (IN 100-BASIS POINTS AT QUARTERLY RATE)

Parameter	Germany		France		U.K.	
μ_1	2.407		1.319		0.216	
	(1.132)		(1.164)		(0.845)	
	0.542		−0.249		−0.282	
	(0.061)		(0.118)		(0.090)	
μ_2	−1.164		−3.042		−2.407	
	(1.178)		(1.163)		(1.057)	
	1.171		−1.228		−1.425	
	(0.086)		(0.238)		(0.174)	
p_{11}	0.972		0.916		0.983	
	(0.030)		(0.050)		(0.019)	
p_{22}	0.951		0.889		0.969	
	(0.039)		(0.071)		(0.044)	
Ω_1	36.553	0.269	34.423	−2.157	31.306	−1.950
	(9.639)	(0.358)	(8.819)	(0.733)	(6.702)	(0.577)
	0.269	0.095	−2.157	0.365	−1.950	0.342
	(0.358)	(0.026)	(0.733)	(0.096)	(0.577)	(0.073)
Ω_2	36.222	−0.920	25.068	−2.036	15.282	−0.335
	(9.816)	(0.528)	(7.586)	(0.974)	(5.733)	(0.680)
	−0.920	0.195	−2.036	0.603	−0.335	0.406
	(0.528)	(0.052)	(0.974)	(0.202)	(0.680)	(0.153)

Note: Standard errors are in parentheses.

ance-covariance matrices Ω_1 and Ω_2, whose properties are governed by the behavior of the specification error u_t in (15).

The unrestricted version of this model is thus a simple vector generalization of the process in Section I:

$$(19) \qquad \mathbf{y}_t | s_t \sim N(\mu_{s_t}, \Omega_{s_t}).$$

Our objective now is to maximize[13]

$$(20) \quad \log p(\mathbf{y}_1,\ldots,\mathbf{y}_T;\theta) - (\nu/2) \cdot [\mu_1'\Omega_1^{-1}\mu_1] - (\nu/2)\cdot[\mu_2'\Omega_2^{-1}\mu_2] - \alpha\log|\Omega_1| - \alpha\log|\Omega_2| - 0.5\omega_1^{ee} - 0.1\omega_1^{ii} - 0.5\omega_2^{ee} - 0.1\omega_2^{ii},$$

where ω_2^{ee}, for example, denotes the (1,1) element of Ω_2^{-1}.

We then fit the unrestricted bivariate model (19) to the exchange rate data along with interest rates. The series used for the latter were the average of bid and asked prices on three-month Eurocurrency rates (quarterly rates, in 100-basis points) as of the close of the London market on the last day of the quarter.[14]

The parameter estimates associated with the highest value for (20) are reported in Table 8 along with asymptotic standard errors. Figures 5 through 7 plot the data and imputed change points, with the two means for the interest differential shown as horizontal dashed lines.

[13] The prior $\Lambda = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.1 \end{bmatrix}$ was used in (12) to weight the variance of the exchange rate innovations to five and the variance of interest rate innovations to one. A different scale variable is appropriate since exchange rates are considerably more variable than interest rate differentials.

[14] These are from the data banks of DRI (called WGD03A, WGD03B, FRD03A, FRD03B, UKD03A, and UKD03B). Data were converted from annual to quarterly rates as $100\cdot\{[1+(i/100)]^{0.25}-1\}$, with i the average of the bid and asked returns.

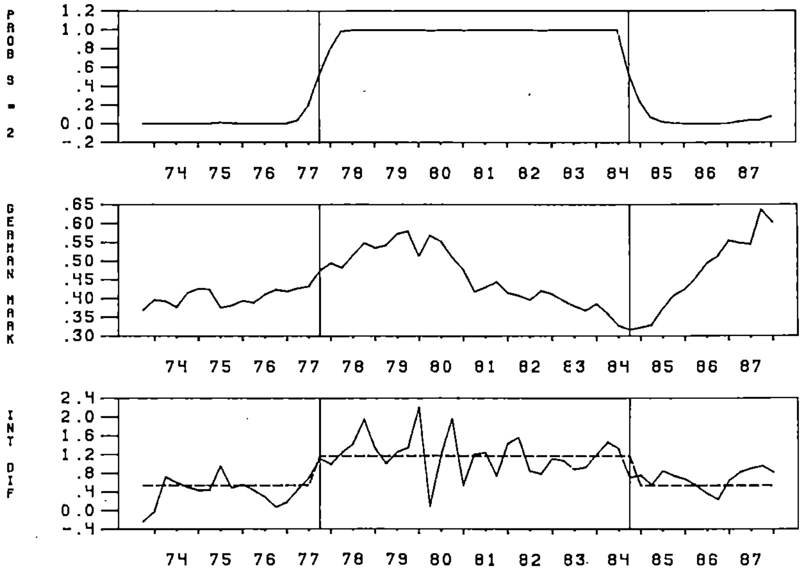


Figure 5

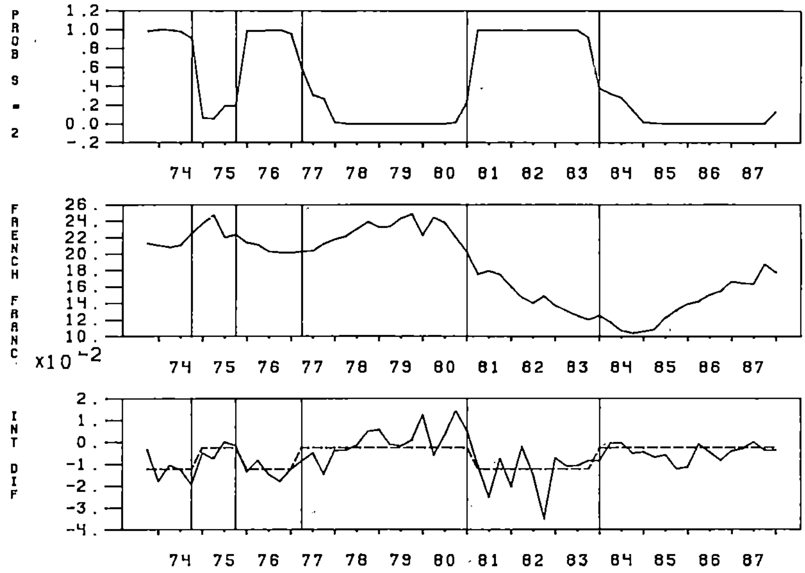


Figure 6

It is difficult to find much support for the hypothesis of uncovered interest parity in these results. Germany is the only country for which the segments identified by the bivariate system (the top panel of Figures 5–7) at all resemble those identified by the univariate process for exchange rates (Figures 2–4), and here the interest differential moves in the opposite direction from that predicted by the theory—the period when the mark was falling was a period when U.S. interest rates were unusually *high* relative

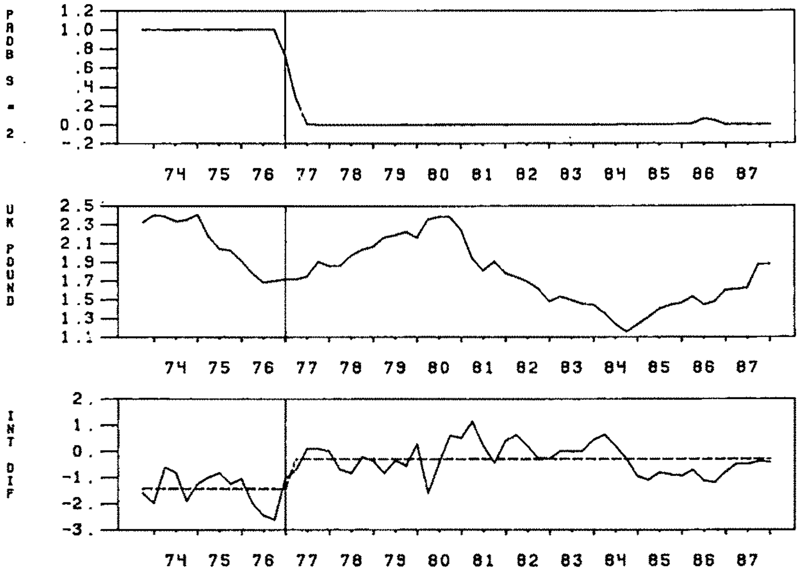


FIGURE 7

to Germany. The interest differential is 62.9 basis points higher in state 2 than in state 1 according to the estimates in Table 8 (1.171 − 0.542 = 0.629), rather than 400 basis points lower as predicted from the univariate estimates in row (3a) of Table 7.

There are some interesting statistics that help to reveal the inconsistencies between the univariate model of exchange rates and a bivariate model that imposes uncovered interest parity. From the standard errors in row (1a) of Table 7 we can construct 95 percent confidence intervals from the univariate estimates for the predicted level of the interest differential in state 1. That is, we can construct confidence intervals for $p_{11}\mu_1 + (1 - p_{11})\mu_2$, which is the predicted value for $i^{US} - i^{WG}$ if we are in state 1. These confidence intervals never overlap with the 95 percent confidence interval for the interest differential from the bivariate estimation for state 1 (row (1b), in Table 7). This is a conservative test in that the marginal significance level is strictly less than 0.05. This is because if the true parameter vector were in the gap between the two confidence intervals, two events (either of which alone has probability less than 0.05 of having occurred) would have to have *both* occurred. Even if the events were perfectly correlated, the probability of both occurring together could be no greater than 0.05, and in general it must be less than 0.05.

Row (3a) of Table 7 gives the change in the interest differential in moving from state 2 to state 1 predicted by the univariate estimates. Row (3b) compares these with the estimates of the actual change in the interest differential, $\mu_1(2) - \mu_2(2)$, as inferred from the bivariate system, where the subscript refers to the state and the "(2)" indicates the second element of the vector $\boldsymbol{\mu}$. In no case do the confidence intervals overlap. This offers evidence against not only the hypothesis of interest parity, but also of a constant risk premium.

The above calculations assumed that, unlike the econometrician, agents knew the state of the process s_t governing the most recent observation on exchange rates $(e_t - e_{t-1})$ with certainty at date t. Charles Engel (1985) and Karen Lewis (1989), for example, discussed models of the exchange rate in which individuals do not know the current monetary policy regime and learn about it gradually through Bayesian inference. Our results change little if we postulate that agents are learning about the state

s_t in the same way as the econometrician. The real-time forecast of the exchange rate change in this case would not be (17) but rather (14):

$$E[y_{t+1}|y_t, y_{t-1},\dots,y_1;\hat{\boldsymbol{\theta}}] = \hat{\mu}_2 + \{\hat{\rho} + (-1+\hat{p}_{11}+\hat{p}_{22}) \cdot [p(s_t=1|y_1,\dots,y_t;\hat{\boldsymbol{\theta}}) - \hat{\rho}]\}\cdot\{\hat{\mu}_1 - \hat{\mu}_2\},$$

which collapses to (17) in the special case when the econometrician has no uncertainty about the state ($p(s_t=1|y_1,\dots,y_t;\hat{\boldsymbol{\theta}}) = 0$ or 1). Equation (18) then becomes

$$(21)\quad i_t^{US} - i_t^{WG} = \hat{\mu}_2 + \{\hat{\rho} + (-1+\hat{p}_{11}+\hat{p}_{22}) \cdot [p(s_t=1|y_1,\dots,y_t;\hat{\boldsymbol{\theta}}) - \hat{\rho}]\} \cdot \{\hat{\mu}_1 - \hat{\mu}_2\} + u_t.$$

Hamilton (1988a) showed how equation (21) could be estimated jointly with the process for exchange rates. Here we settle for a more modest descriptive statistic, obtained from the regression of the interest differential on the output of the filter from the univariate estimator

$$i_t^{US} - i_t^{WG} = \beta_0 + \beta_1[p(s_t=1|y_1^{WG}, y_2^{WG},\dots,y_t^{WG};\hat{\boldsymbol{\theta}})] + u_t.$$

This OLS regression has an R^2 of 0.01 for all three currencies, which we take as convincing evidence that uncovered interest parity can not explain much of the movements in interest differentials.

Thus neither the assumption that markets know the regime with certainty nor the assumption that they are learning about it through the rule $p(s_t=1|y_1,\dots,y_t;\hat{\boldsymbol{\theta}})$ offers a very appealing account of time variation in cross-country interest differentials.

V. Conclusion

Movements in the dollar appear to be characterized by long swings. We have presented a formal statistical model of what it means for the dollar to follow a pattern of long swings, and we find that the model fits the data well. We conclude that the phenomenon of long swings deserves more attention from exchange rate theoreticians.

Can we offer an explanation of these exchange rate and interest rate movements? Dornbusch (1986, 1987), Bernard Dumas (1987), Stephan Schulmeister (1987), and Betty Daniel (1989) have suggested models that allow persistence in movements in the exchange rate. Robert Flood and Peter Garber (1983) have discussed models in which anticipated future events affect the current exchange rate, generating nonlinear behavior of the exchange rate akin to that described here. Hsieh (1988) has described a model in which monetary policy stochastically shifts between two regimes. Kaminsky (1988) has generated a simple model that also leads to nominal exchange rate movements of the type we describe in Section I, though empirically identifying the particular fundamentals that have shifted in the way postulated by her model poses a challenge for future research. Jeffrey Frankel and Kenneth Froot (1988) have described a model in which the behavior of irrational "chartists" interacts with rational agents to produce potentially long movements in one direction in the dollar and a failure of uncovered interest parity. These models seem able to account for some, but not all, of the empirical regularities uncovered here.

A model that allowed only rational investors would need to explain the pattern of the dollar and be able to generate risk premia that varied enough over time to explain the pattern of interest differentials. This is an imposing task (although see Cumby, 1988).

Earlier researchers found little evidence of linear serial dependence in exchange rate changes, supporting the conclusion that the exchange rate follows a random walk. We reproduce this result but nevertheless find compelling evidence of nonlinear serial dependence in the data characteristic of long swings. Our evidence indicates that movements in the dollar in one direction persist over long periods of time. Furthermore, interest differentials do not seem to take into account how long these movements are. Our estimation method provides a natural way of parameterizing the "peso problem," yet we still reject the uncovered interest parity hypothesis. In the absence of a plausible story about foreign exchange risk premia, we conclude that there are long swings in the dollar and that markets do not know it.

APPENDIX

Treating each currency separately, we began from an initial starting value for $\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, p_{11}^{(0)}, p_{22}^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)})'$. We then iterated on equations (2), (9), (10), (7), and (8) until the largest element of $\hat{\theta}^{(\ell+1)} - \hat{\theta}^{(\ell)}$ was less than 1×10^{-8} in absolute value. For the Bayesian parameter ν that appears in (9), we specified $\nu = 0.1$, which roughly corresponds to proceeding as if we had observed one-tenth of an observation drawn from each regime that took on the value zero. We further specified $\alpha = 0.1$ and $\beta = 0.5$, as if we had two-tenths of an observation from each regime whose sum of squared deviations from the population mean of that regime was $\beta / \alpha = 5$.

For each currency we employed several hundred different starting values $\theta^{(0)}$. These starting values all led to a single unique solution to the normal equations in the case of France. However, two local maxima were found for Britain and four for Germany. The maximum likelihood estimates reported in Table 1 are those that achieved the highest value of the objective function (11). When a diffuse Bayesian prior is used ($\nu = \alpha = \beta = 0$) and iteration is begun from the starting values in Table 1, the parameter estimates are changed very little. It further appears that, apart from the singularities, these correspond to the largest bounded local maxima of the raw likelihood function (4). Thus, use of the prior has essentially no consequences for any of the tests of conclusions reported in this paper. By contrast, other local maxima of (4) or (11) exhibit large changes in parameter estimates for slight changes in the priors. Hamilton (1988b) has argued that such a finding should be construed as an additional factor supporting selection of the global maxima reported in Table 1 for the U.K. and Germany.

Our bivariate analysis found four local maxima for France and Germany and six for the U.K.

REFERENCES

Blanchard, Olivier and Summers, Lawrence H., "Perspectives on High World Real Interest Rates," *Brookings Papers on Economic Activity*, 1984, *2*, 273–334.

Borensztein, Eduardo, "Alternative Hypotheses About the Excess Return on Dollar Accounts, 1980–1984," *IMF Staff Papers*, March 1987, *34*, 29–59.

Brock, William, Dechert, W. Davis and Scheinkman, José, "A Test for Independence Based on the Correlation Dimension," mimeo., University of Wisconsin, Madison, 1987.

Cochrane, John H., "How Big Is the Random Walk in GNP?" *Journal of Political Economy*, October 1988, *96*, 893–920.

Cumby, Robert, "Is It Real? Explaining Deviation from Uncovered Interest Parity," *Journal of Monetary Economics*, September 1988, *22*, 279–99.

______ **and Obstfeld, Maurice,** "International Interest Rate and Price Level Linkages Under Flexible Exchange Rates: A Review of Recent Evidence;" in John F. O. Bilson and Richard C. Marston, eds., *Exchange Rate Theory and Practice*, Chicago: University of Chicago Press, 1984.

Daniel, Betty C., "One-Sided Uncertainty About Future Fiscal Policy," *Journal of Money, Credit, and Banking*, May 1989, *21*, 176–89.

Diebold, Francis X. and Nason, James M., "Nonparametric Exchange Rate Prediction?," *Journal of International Economics*, forthcoming.

Dooley, Michael P. and Isard, Peter, "The Appreciation of the Dollar: An Analysis of the Safe-Haven Hypothesis," IMF working paper, 1985.

Dornbusch, Rudiger, "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, *84*, 1161–76.

______, "Comment on Shafer and Loopesko," *Brookings Papers on Economic Activity*, 1983, *1*, 79–85.

______, "Flexible Exchange Rates and Excess Capital Mobility," *Brookings Papers on Economic Activity*, 1986, *1*, 209–26.

______, "Exchange Rate Economics: 1986," *Economic Journal*, March 1987, *97*, 1–18.

Dumas, Bernard, "Pricing Physical Assets Internationally: A Nonlinear Heteroskedastic Process for Equilibrium Real Exchange Rates," mimeo., University of Pennsylvania, 1987.

Engel, Charles, "Reliability of Policy Announcements and the Effects of Monetary Policy," *European Economic Review*, November 1985, *29*, 137–55.

Evans, George W., "A Test for Speculative Bubbles in the Sterling-Dollar Exchange

Rate: 1981–84," *American Economic Review*, September 1986, *76*, 621–36.

Everitt, B. S. and Hand, D. J., *Finite Mixture Distributions*, London: Chapman and Hall, 1981.

Feldstein, Martin, "The Budget Deficit and the Dollar," in Stanley Fischer, ed., *NBER Macroeconomics Annual 1986*, Cambridge, MA: MIT Press, 1986.

______, "Let the Market Decide," *The Economist*, December 3–9, 1988, *309*, 21–24.

Flood, Robert and Garber, Peter, "A Model of Stochastic Process Switching," *Econometrica*, May 1983, *51*, 537–51.

Frankel, Jeffrey, "The Dazzling Dollar," *Brookings Papers on Economic Activity*, 1985, *1*, 199–218.

______, "International Capital Flows and Domestic Economic Policies," in Martin Feldstein, ed., *The United States and the World Economy*, Chicago: University of Chicago Press, 1988.

______ **and Froot, Kenneth,** "Using Survey Data to Test Standard Propositions Regarding Exchange Rate Expectations," *American Economic Review*, March 1987, *77*, 133–53.

______ **and** ______, "Chartists, Fundamentalists, and the Demand for Dollars," in Tony Courakis and Mark Taylor, eds., *Policy Issues for Interdependent Economies*, London: Macmillan, 1988.

Gallant, A. Ronald, *Nonlinear Statistical Models*, New York: Wiley & Sons, 1987.

Hamilton, James D., (1988a), "Rational-Expectations Econometric Analysis of Changes in Regime: An Investigation of the Term Structure of Interest Rates," *Journal of Economic Dynamics and Control*, June/September 1988, *12*, 385–423.

______, (1988b), "A Pseudo-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions," mimeo., University of Virginia, 1988.

______, (1989a), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, March 1989, *57*, 357–84.

______, (1989b), "Specification Testing in Markov-Switching Time Series Models," mimeo., University of Virginia, 1989.

______, "Analysis of Time-Series Subject to Changes in Regime," *Journal of Econometrics*, forthcoming.

Hodrick, Robert J., *The Empirical Evidence on the Efficiency of Forward and Futures Foreign Exchange Markets*, New York: Harwood Academic Publishers, 1987.

______, **and Srivastava, Sanjay,** "An Investigation of Risk and Return in Forward Foreign Exchange," *Journal of International Money and Finance*, April 1984, *3*, 249–65.

Hsieh, David, "A Nonlinear Stochastic Rational Expectations Model of Exchange Rates," mimeo., University of Chicago, 1988.

______, "Testing for Nonlinear Dependence in Daily Foreign Exchange Rates," *Journal of Business*, July 1989, *62*, 339–68.

Kaminsky, Graciela, "The Peso Problem and the Behavior of the Exchange Rate: The Dollar/Pound Exchange Rate, 1976–1987," mimeo., U.C.-San Diego, 1988.

Krasker, William S., "The 'Peso Problem' in Testing the Efficiency of Forward Exchange Markets," *Journal of Monetary Economics*, April 1980, *6*, 269–76.

Lee, Lung-Fei and Chesher, Andrew, "Specification Testing When Score Statistics Are Identically Zero," *Journal of Econometrics*, March 1986, *31*, 121–49.

Lewis, Karen K., "Can Learning Affect Exchange Rate Behavior? The Case of the Dollar in the Early 1980s," *Journal of Monetary Economics*, January 1989, *23*, 79–100.

Meese, Richard and Rogoff, Kenneth, (1983a), "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?" *Journal of International Economics*, February 1983, *14*, 3–24.

______ **and** ______, (1983b), "The Out-of-Sample Failure of Empirical Exchange Rate Models: Sampling Error or Misspecification?," in Jacob Frenkel, ed., *Exchange Rates and International Macroeconomics*, Chicago: University of Chicago Press, 1983.

______ **and Singleton, Kenneth,** "On Unit Roots and the Empirical Modeling of Exchange Rates," *Journal of Finance*, September 1982, *37*, 1029–35.

Mussa, Michael, "Empirical Regularities in the Behavior of Exchange Rates and Theories of the Foreign Exchange Market," *Carnegie-Rochester Conference Series on Public Policy*, 1979, *11*, 9–57.

______, "Nominal Exchange Rate Regimes and the Behavior of Real Exchange Rates: Evidence and Implications," *Carnegie-Rochester Conference Series on Public Policy*, 1986, *25*, 117–214.

Newey, Whitney K., "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, September 1985, *53*, 1047–70.

Schulmeister, Stephan, "An Essay on Exchange Rate Dynamics," mimeo., Austrian Institute of Economic Research, 1987.

Tauchen, George, "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics*, October/November 1985, *30*, 415–43.

Watson, Mark W. and Engle, Robert F., "Testing for Regression Coefficient Stability with a Stationary AR(1) Alternative," *Review of Economics and Statistics*, May 1985, *67*, 341–46.

White, Halbert, "Specification Testing in Dynamic Models," in Truman F. Bewley, ed., *Advances in Econometrics, Fifth World Congress*, Vol. I, Cambridge: Cambridge University Press, 1987.

Testing the Rationality of Price Forecasts: New Evidence from Panel Data

By MICHAEL P. KEANE AND DAVID E. RUNKLE*

This paper tests the rationality of individual price forecasts in a panel of professional forecasters. Here, unlike in most previous studies, rationality is not rejected. The results here differ because (1) using individual forecasts avoids aggregation bias, (2) comparison of forecasts to initial data avoids bias due to data revision, (3) the professional forecasters have economic incentives to state their expectations accurately, (4) a new covariance matrix estimator consistent when forecast errors are correlated across individuals is used. (JEL 132)

How people form their expectations of future economic events has been an important issue in macroeconomics for many years. Businessmen's expectations play a central role in the business-cycle theories of both Arthur Pigou (1927) and John Maynard Keynes (1936). Since the way in which expectations are formed has important implications for economic behavior, many economists have used survey data to test hypotheses about expectation formation. In particular, several researchers have tested the hypothesis that expectations are rational in John Muth's sense (Muth, 1961). In this paper, we provide a further test of the rational expectations hypothesis. Specifically, we test the rationality of price-level forecasts. We use price-forecast data from the ASA-NBER survey of professional forecasters, initiated by Victor Zarnowitz (1969, 1974, 1984, 1985). Zarnowitz, like most other researchers in this literature, finds that price forecasts are not rational.[1] However, using different statistical methods, we find strong evidence that price forecasts are rational.

We believe that our further testing of price-forecast rationality is warranted because of severe problems in almost all existing tests, which suffer from one or more of the following four flaws. First, some use average survey response data rather than individual data. This can bias tests in two ways. It can lead to false rejection of rational expectations because average forecasts that are conditional on *different* information sets are not rational forecasts conditional on any *particular* information set. Further,

*Keane, University of Minnesota, Minneapolis, MN 55455, and Federal Reserve Bank of Minneapolis, Minneapolis, MN 55480; Runkle, Federal Reserve Bank of Minneapolis and University of Minnesota. Keane would like to thank the Alfred P. Sloan Foundation and the Federal Reserve Bank of Minneapolis for their support. Both authors would like to thank Lars Hansen, Whitney Newey, Christopher Sims, Victor Zarnowitz, two anonymous referees, and seminar participants at the University of Minnesota and the 1987 Winter Meetings of the Econometric Society for helpful suggestions; Daniel Chin and Ruth Judson for research assistance; and Patricia McKernon for editorial assistance. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

The authors' previously published paper, Keane and Runkle (1989), also deals with the subject of price forecast rationality. The main object of that paper is to explain for a general audience the basic issues involved in testing rationality. It presents several of the rationality test results presented here. However, that paper does not describe the statistical methods we develop to deal with aggregate shocks and moving-average errors, or discuss the issues of aggregation bias or the problems involved in measuring expectations. Nor does it contain the empirical results on the importance of aggregate shocks and moving-average errors, the results on the importance of data revisions, or the results on private information.

[1] Zarnowitz (1985) found that the forecasts of other variables, besides prices, were rational. We explain his finding of price-forecast irrationality. Lovell (1986) reviews the literature on tests of rational expectations.

it can lead to false acceptance of rational expectations by masking systematic individual bias that may be randomly distributed in the population. Second, many tests fail to deal properly with the pervasive problem of systematic data revision. Tests of forecast rationality depend upon correct assumptions about what the forecasters tried to predict and what they knew when they made their predictions. Much work has implicitly tested whether forecasters rationally forecast revised data conditional on other revised data, none of which was available until long after the forecasts were made. Third, most studies of forecast rationality use predictions from individuals who are not professional forecasters; these people have few economic incentives to report their expectations precisely. Finally, many studies that use micro data fail to account properly for the covariance structure of the forecast errors. This failure can take two forms. First, some studies assume that forecast errors must be white noise. In fact, lags in the availability of relevant data can produce serially correlated errors even when agents are rational. Second, most studies fail to account for the fact that shocks to the aggregate economy produce forecast errors that are correlated across individuals. In either case, improperly assuming independent, identically distributed errors can produce severely biased results.

In this paper we avoid these problems. We use the individual data on quarterly GNP deflator forecasts from the ASA-NBER survey of forecasters to test the rational expectations hypothesis. We treat the data as a panel to avoid aggregation biases, and we reconstruct the actual information sets available to agents when they made their forecasts. We compare these forecasts only to the GNP deflator data announced 45 days after the end of the quarter, before systematic data revisions. The survey data include only forecasts from professional forecasters, who have an economic incentive to be accurate. Because these professionals report to the survey the same forecasts that they sell on the market, their survey responses provide a reasonably accurate measure of their expectations. Thus, these data are less subject to the criticism made by opponents of survey forecast rationality tests that the respondents had nothing to lose if they made bad forecasts.[2] Finally, we develop a new covariance matrix that is consistent both in the presence of aggregate shocks and in the presence of serial correlation resulting from delays in data availability. When we use all these procedures, we find that the rational expectations hypothesis cannot be rejected, and we demonstrate how the hypothesis is rejected when some of these procedures are not followed. We also report some findings that extend beyond the existing literature. We test the hypothesis that differences in individual forecasts are due solely to asymmetric information and cannot be explained by any publicly available information. To our knowledge, this strong implication of the rational expectations hypothesis has not been tested previously. We find that it cannot be rejected. Finally, we find that, because of lagged data availability, forecast errors are MA(1) despite the rationality of the forecasters. This error structure implies that the class of rational expectations models in which output varies only because of monetary surprises can produce persistent movement in output.

The outline of the paper is as follows. Section I addresses the argument, made by some economists, that the rational expecta-

[2] Note that some authors (for example, Cukierman, 1986) have called this inaccuracy "measurement error." But what these authors are discussing is different from the usual interpretations of measurement error. An example will make quite clear what they call "measurement error." Suppose someone calls Keane and asks for his forecast of the three-month T-bill rate for the next quarter. He is busy writing a paper for a conference—the activity for which he receives monetary reward—that is due in three days. Quickly, Keane tells the caller 8 percent. While reading the *Wall Street Journal* later in the day, Keane sees that the forward rate on three-month T-bills is 9 percent. He does not run out to buy bonds in the expectation that rates will fall to 8 percent because, when he thinks about it, 9 percent seems reasonable. Thus, 8 percent is an erroneous measure of his true expectation because he does not act in the market as if that were his expectation. But that is not measurement error as it is usually understood, so we prefer to call the problem "lack of an economic incentive to accurately state expectations."

tions hypothesis cannot be tested using survey data. Section II reviews the literature and describes the problems with previous research that have motivated our work. Section III discusses our econometric methods, including a new covariance matrix estimator for panel data that is consistent in the presence of aggregate shocks. Section IV describes the ASA-NBER survey data and discusses the complex data timing and revision issues. Section V presents our empirical results. Section VI concludes.

I. On the Use of Survey Data to Test Theories of Expectation Formation

Some economists have questioned whether it is valid to use survey data on agents' expectations to test among various mechanisms of expectation formation. For example, Edward Prescott (1977, p. 30) has argued that "surveys cannot be used to test the rational expectations hypothesis. One can only test if some [economic] theory, whether it incorporates rational expectations or, for ... [that] matter, irrational expectations, is or is not consistent with observations." Others, such as Zarnowitz (1984, p. 15), have argued that "it is not good 'positive economics' to dismiss it [evidence from survey data] on the ground that only theories, not their assumptions, can be tested."

Economists do agree that the ability of economic models to explain behavior depends on the assumed expectation-formation mechanism. For example, the intertemporal substitution model proposed by Robert Lucas and Leonard Rapping (1969) provides a far better explanation of employment fluctuations if expectations are "adaptive" rather than rational. Since any test of a model involves a joint test of the behavioral equations and the assumed expectation-formation mechanism, researchers face an identification problem. First, if the model is rejected, we do not know whether the behavioral equations or the expectations mechanism is being rejected. Second, if two different models (different in both the behavioral equations and expectation-formation mechanism) explain data equally well, we cannot identify the proper behavioral model without knowing the expectation-formation mechanism.[3]

Many economists have assumed the validity of the rational expectations hypothesis and regarded joint tests of behavioral equations and this particular expectation-formation mechanism as identified tests of behavioral equations. This willingness to assume the validity of Muth's rational expectations hypothesis perhaps results from a widespread view that forming expectations rationally is simply the logical consequence of optimizing behavior. Yet there are many reasons why the rational expectations hypothesis *does not* follow directly from the assumption of optimizing behavior. For example, agents' expectations in stochastic equilibrium may be rational in Muth's sense, while the expectations of Bayesian learners in a nonstationary environment are not (see John Caskey, 1985). Or, perhaps, information necessary to form expectations on the basis of the true economic model may be impossible or too costly to obtain (see Kenneth Arrow, 1978). In this paper we adopt Zarnowitz's view that the hypothesis of rational expectations should be tested and not simply assumed valid. For further discussion of the philosophical issues, we refer the reader to Michael Lovell (1986) or Zarnowitz (1984).

II. A Critique of the Literature

Given the large number of empirical tests of the rational expectations hypothesis in the literature, one might wonder whether yet another test is necessary. We contend that almost all existing tests are either incorrect or inadequate for four reasons. First, most tests use sample mean forecasts rather than individual forecasts. Second, respondents in most surveys have little incentive to make accurate forecasts. Third, most tests compare survey forecasts to revised rather than initial data. Fourth, many tests are based on incorrect assumptions about the

[3] This is the well-known problem of observational equivalence (see Sargent, 1976).

covariance structure of forecast errors. We discuss these problems with the existing literature in this section.

Many tests of the rational expectations hypothesis that use survey data examine the rationality of the sample mean or "consensus" forecast constructed from surveys of individual forecasters. These include most tests using the Livingston data,[4] as well as the work by Jonathan Leonard (1982) on the Endicott survey of employers' wage expectations, that by Frank de Leeuw and Michael McKelvey (1981) on the price expectations of business firms, that by Jeffrey Frankel and Kenneth Froot (1987) on various surveys of exchange rate expectations, and that by Benjamin Friedman (1980) on the Goldsmith-Nagen survey of interest rate expectations.

There are two problems with using consensus forecasts to test rationality. First, doing so causes serious specification bias. If forecasters are rational, their forecasts will differ only because of differences in their information sets.[5] The mean of many individual rational forecasts, each conditional on a private information set, is not itself a rational forecast conditional on any particular information set (see Stephen Figlewski and Paul Wachtel, 1983). This seemingly minor issue can produce severe bias. We discuss this problem further in Section III, Part B.

A second problem with using consensus forecasts is that this approach can mask individual deviations from rationality. Albert Hirsch and Michael Lovell (1969), looking at data from individual firms in the Commerce Department Manufacturers' Inventory and Sales Expectations survey, found (p. 71) that some firms are consistently optimistic about future sales while others are consistently pessimistic. Averaging expectations, however, can cancel these biases across firms so that industry mean expectations show no bias. Muth (1985) looked at anticipated production for individual Pittsburgh steel firms and found the same phenomenon.

For both of these reasons we argue that researchers must use *individual* data in order to test hypotheses about how people form expectations. These data can be used to test rationality either on an individual-by-individual basis or by running pooled time-series cross-section regressions. Both individual and pooled regressions are represented in the work of Hirsch and Lovell (1969), Muth (1985), Figlewski and Wachtel (1981), Thomas Urich and Paul Wachtel (1984), de Leeuw and McKelvey (1981), and Zarnowitz (1984, 1985).

A second, equally severe problem with most tests of forecast rationality is that survey data on expectations do not necessarily reflect the true expectations of the forecasters. But in order to test the rational expectations hypothesis, one needs data that can be reasonably assumed to reflect the forecasters' expectations. As de Leeuw and McKelvey (1984) found, this is certainly not the case with surveys, such as the BEA data on sales price expectations of individual firms, that ask individuals or firms to give "rough estimates" of expected future quantities. As mentioned earlier, Hirsch and Lovell (1969) and Muth (1985) found evidence against rationality on the firm level. However, the problem found by de Leeuw and McKelvey may well plague the Hirsch-Lovell and Muth studies as well. As we discussed in the introduction, it is reasonable to assume that this problem is limited when dealing with the forecasts of professional forecasters who receive a monetary reward for producing accurate forecasts and who report to the survey the same forecasts they sell on the market. This is not true of the Livingston price forecast data, because the economists polled by Livingston were not all professional forecasters. Hence we cannot be sure that the rejection of rationality found by Figlewski and Wachtel (1981), who ran a pooled regression using the

[4] See Brown and Maital (1981), Carlson (1977), Gramlich (1983), Mullineaux (1978), Pearce (1979), Pesando (1975), and Rich (1987).

[5] This is a very important point. If each forecaster had exactly the same information, all forecasts would have to be the same, if they were rational. Forecasters can make different rational forecasts if they each have some different private information. We discuss this issue more fully in Section III, Part C.

Livingston data, is not the result of improperly measured expectations.

Since 1968 Victor Zarnowitz has worked with the NBER and the American Statistical Association on the ASA-NBER Survey of Forecasts by Economic Statisticians.[6] This survey is well suited for testing because it is limited to professional producers of quarterly forecasts of GNP and its major components and other economic indicators for each of several quarters ahead. Since respondents report forecasts that they produce professionally, the problem of inaccurate reporting of expectations is probably small. Interestingly, the ASA-NBER also conducted a few annual surveys, in the early years of the period covered, of a much larger group of economists of whom three-quarters are "occasional" rather than "professional" forecasters. In this survey, with method similar to that of Livingston, Zarnowitz (1969) finds that "a number of the occasional forecasters submitted extreme and rather unreasonable predictions." This finding adds to our concern about inaccuracies due to lack of proper economic incentives in the Livingston data.

Using the ASA-NBER survey, Zarnowitz (1984, 1985) has performed tests of the rationality of the respondents' forecasts of inflation and other macroeconomic variables. Using both survey means and pooled data, he rejects the rational expectations hypothesis for inflation forecasts. We consider these results questionable, however, because Zarnowitz uses revised rather than initial price data in his tests. Using revised data does not necessarily invalidate tests of unbiasedness, but their use is suspect in this case because of the nature of the data revisions.[7] We show in Section V, Part B that these data revisions introduce a systematic bias that may invalidate Zarnowitz's tests of unbiasedness.

A final problem with much of the existing literature is incorrect assumptions about the covariance of forecast errors across forecasters. When they pool individual data, Figlewski and Wachtel (1981) assume that forecast errors are independent across forecasters.[8] Since aggregate shocks affect the price level, this is certainly not true. Falsely assuming independent errors creates a severe downward bias in estimated standard errors, tending to cause false rejection of the rational expectations hypothesis. In Section III, Part D we discuss how to account correctly for the effects of aggregate shocks on inference in tests of forecast rationality. In Section V, Part A we show the empirical importance of these effects.

III. Econometric Issues

A. *Definitions*

Expectations are rational in Muth's sense (Muth, 1961) if they are equal to mathematical expectations conditional on the set of all information relevant for forecasting. For an individual forecaster, we can express this relationship as

$$(1)\qquad {}_tP_{i,t+k} = E(P_{t+k} \mid I_{i,t}),$$

where P_{t+k} is the realized value of the time series P at time $t+k$, ${}_tP_{i,t+k}$ is a k-step-ahead prediction of P made at time t by forecaster i, $I_{i,t}$ is the information available at time t to forecaster i, and E is the mathematical expectation operator. This is equivalent to the statement: $E(\varepsilon^i_{t,k} \mid I_{i,t}) = 0$, where $\varepsilon^i_{t,k} = P_{t+k} - {}_tP_{i,t+k}$. This statement can be broken down into the separate hypotheses that forecasts are unbiased and efficient.

For an individual forecaster, a test of rationality can be performed by running the

[6] For a description of the survey, see Zarnowitz (1969, 1974).

[7] In Section IV we discuss the circumstances under which it is appropriate to use revised data to test for unbiasedness.

[8] Zarnowitz argues, correctly, that it is invalid to pool individual data while assuming independent forecast errors within each cross section if there are aggregate shocks.

regression

$$(2)\quad P_{t+k} = \alpha_0 + \alpha_1 {}_tP_{i,t+k} + \alpha_2 X_{i,t} + \varepsilon^i_{t,k},$$

where $X_{i,t}$ is any variable in forecaster i's information set at time t. Unbiasedness requires that, in a regression without $X_{i,t}$ variables, the coefficients in equation (2) may be restricted to $\alpha_0 = 0$ and $\alpha_1 = 1$. Efficiency requires that any variable known at time t or before be orthogonal to $\varepsilon^i_{t,k}$; that is, $\alpha_2 = 0$ for any $X_{i,t} \in I_{i,t}$.[9]

B. *The Aggregation Problem and the Advantages of Panel Data*

As we discussed in Section II, using survey means to test the rational expectations hypothesis leads to serious specification error. This bias can be illustrated by comparing the results of three different estimation methods. If we were to run three separate tests of unbiasedness using equation (2), first doing a separate regression for each individual and taking the mean of the estimated coefficients, second using the pooled data, and third using survey means, we could call the α_i estimates $\hat{\alpha}_{1i}$, $\hat{\alpha}_{1p}$, and $\hat{\alpha}_{1m}$, respectively, and we would have

$$(3)\qquad \hat{\alpha}_{1i} = \frac{1}{N}\sum_{i=1}^{N} \frac{\Sigma_{t=1}^{T} P_{t+k}\,{}_tP_{i,t+k}}{\Sigma_{t=1}^{T}\,{}_tP^2_{i,t+k}}$$

$$(4)\qquad \hat{\alpha}_{1p} = \frac{\Sigma_{t=1}^{T}\Sigma_{i=1}^{N} P_{t+k}\,{}_tP_{i,t+k}}{\Sigma_{t=1}^{T}\Sigma_{i=1}^{N}\,{}_tP^2_{i,t+k}}$$

$$(5)\qquad \hat{\alpha}_{1m} = \frac{\Sigma_{t=1}^{T} P_{t+k}\left(\Sigma_{i=1}^{N}\,{}_tP_{i,t+k}\right)}{\Sigma_{t=1}^{T}\left(\Sigma_{i=1}^{N}\,{}_tP^2_{i,t+k}\right)},$$

where i indexes individuals from 1 to N. We see that $\text{plim}_T\,\hat{\alpha}_{1i} = \text{plim}_T\,\hat{\alpha}_{1p}$ but that

$$(6)\quad \frac{\hat{\alpha}_{1p}}{\hat{\alpha}_{1m}} = \frac{\text{Var}\left({}_t\bar{P}_{i,t+k}\right)}{\text{Var}\left({}_t\bar{P}_{i,t+k}\right) + \text{Var}\left({}_tP_{i,t+k} - {}_t\bar{P}_{i,t+k}\right)},$$

where ${}_t\bar{P}_{i,t+k} \equiv (1/N)\Sigma_{i=1}^{N}\,{}_tP_{i,t+k}$ and the grand mean of ${}_tP_{i,t+k}$ is assumed to be 0. Thus, as long as individual forecasts differ, $\hat{\alpha}_{1m}$ will be biased upward. The importance of this bias can be seen in the work of Urich and Wachtel (1984) who, using 20 individuals and 95 time periods to test the rationality of money supply forecasts, obtained the following estimates:

Estimator	α_0	α_1
Individual Means	−0.13	0.78
Pooled Data	−0.12	0.77
Sample Means	−0.29	1.06

The upward bias in the sample means estimator $\hat{\alpha}_{1m}$ is seen here to be quite severe (about 40 percent). We avoid this aggregation problem by using only individual data in a pooled time-series cross-section regression.

By using panel data to test the rational expectations hypothesis, we avoid the aggregation problem. We also achieve two other important advantages over time-series data. First, we can detect systematic bias of individual forecasters by testing for the presence of individual effects. Second, using panel data increases the degrees of freedom available to test hypotheses, making our tests of rationality more powerful.

It is not immediately clear that using panel data makes tests of rationality more powerful. Since the dependent variable is the same for all individuals, forecast errors will be highly correlated across individuals and the effective number of degrees of freedom will not equal the number of observations minus the number of parameters. OLS standard errors will therefore be severely biased. Other authors who have used panel data (for example, Figlewski and Wachtel, 1981; Urich and Wachtel, 1984; and de Leeuw and McKelvey, 1984) have not adjusted the OLS standard errors. We discuss our

[9] If forecasters have differential costs of over- and underprediction, it could be rational for them to produce biased forecasts. If we were to find that forecasts are biased, it could still be claimed that forecasters were rational if it could be shown that they had such differential costs.

method of consistently estimating the covariance matrix in Section III, Part D.

C. *The Forecasters' Information Sets*

We now address two issues concerning the covariance structure of the errors in (2). First, what is the information structure implied by (2)? Second, what are the effects of aggregate shocks on inference in (2)? Certainly any public information known at time t or before should be orthogonal to $\varepsilon^i_{t,k}$. But an additional restriction requires that any private information known by person i at time t or before also be orthogonal to $\varepsilon^i_{t,k}$. One's own prior forecasts and prior forecast errors are examples of such private information. Further, if forecasts are publicly announced, then other people's prior forecasts and prior forecast errors should be orthogonal to $\varepsilon^i_{t,k}$ as well. Finally, the rational expectations hypothesis implies that forecasts should only differ across individuals if individuals have different private information (i.e., if the conditioning set $I_{i,t}$ differs among forecasters). This implies that no readily available data should be able to explain the differences between individuals' forecasts. This proposition is implied by rational expectations but has not been tested previously.

Unfortunately, merely stating these constraints does not tell us what forecasters should know for a particular survey. Certainly, forecasters should know their own forecasts in this and previous periods, so ${}_tP_{i,t+k-1}$, and ${}_{t-1}P_{i,t+k-1}$ should both be orthogonal to $\varepsilon^i_{t,k}$. The most difficult informational issue is whether P_t itself is known when the forecast ${}_tP_{i,t+k}$ is made. In many surveys, P_t is not released until after forecasters have predicted ${}_tP_{i,t+k}$. In those surveys, neither P_t nor the lagged one-step-ahead forecast error $\varepsilon^i_{t-1,1} \equiv P_t - {}_{t-1}P_{i,t}$ should be orthogonal to $\varepsilon^i_{t,k}$. Therefore, the forecast errors will be MA(k), rather than MA($k-1$), as they would be if the forecasters knew P_t when they made their forecasts.[10] This serial correlation does not refute rationality. However, rationality cannot hold if higher-order serial correlation exists, because such serial correlation would imply that individual forecasters do not learn from their own past errors.

We also note that if we observe ${}_tP_{i,t}$, both it and forecasters' perceptions of their own lagged one-step-ahead forecast errors (${}_tP_{i,t} - {}_{t-1}P_{i,t}$) should be orthogonal to $\varepsilon^i_{t,k}$.[11] Of course, information that forecasters did not know when they predicted ${}_tP_{i,t+k}$ should not be orthogonal to $\varepsilon^i_{t,k}$. That would include, for example, such quantities as P_t, ${}_{t+1}P_{i,t+k}$, ${}_{t+1}P_{i,t+k+1}$, $P_t - {}_{t-1}P_{i,t}$, and $P_t - {}_tP_{i,t}$. Additionally, if the past forecast of another forecaster (${}_{t-1}P_{j,t}$) is publicly announced, then a forecaster's perception of the other forecaster's lagged error (${}_tP_{i,t} - {}_{t-1}P_{j,t}$) should not improve price forecasts.

D. *Aggregate Shocks and Tests of Rationality*

At first glance, (2) seems trivial to estimate. It would seem that we could use OLS and estimate a covariance matrix that is consistent in the presence of serial correlation. However, to use these estimators is to assume that forecast errors are uncorrelated across forecasters. But such correlation is likely because of aggregate shocks to the economy. If the average forecast error is not zero for each period, then OLS will yield inconsistent parameter estimates. Chamberlain (1984) first noted the potential effects of aggregate shocks on rational expectations models estimated with panel data. Chamberlain suggests that one result of aggregate shocks in a rational expectations model is that the sample version of the orthogonality condition $E(\varepsilon^i_{t,k} \mid I_{i,t})$ converges to zero as the number of time peri-

[10] In a different context, Watson (1983) made this observation. Note also that we cannot use a GLS transformation on (2) in this case because the regressors are not strictly exogenous.

[11] ${}_tP_{i,t}$ is the guess that person i made at the end of time t about the value of P_t. Since data on P_t may not be released until after that guess is made, rationality does not imply that ${}_tP_{i,t} = P_t$. Note also that data revisions made after the forecast are not in the forecaster's information set when he makes the prediction, and therefore functions of those revisions cannot be used as independent variables in efficiency tests.

ods increases, but not as the number of individuals increases, if the number of time periods is held fixed. He also notes that most panel data models rely on a large number of individuals to achieve consistency. Thus, in a panel with a large N and a small T, coefficients in rational expectations models may be inconsistent.[12] This problem can be addressed in two ways. One can use either a long panel so that aggregate shocks do not affect consistency or time dummies to eliminate the effects of aggregate shocks. In this study, we use a long panel to achieve consistency.[13]

OLS also yields inconsistent standard errors in the presence of aggregate shocks. In this paper, we present an important innovation: In order to achieve efficiency, we use a covariance matrix estimator that remains consistent in the presence of aggregate shocks.

We now describe the structure of the covariance matrix estimator. Recall that the GMM estimator in a linear model is

$$(7)\quad \hat{\beta}_{GMM}=\left(\frac{X'Z}{NT}\left(\frac{Z'\Omega Z}{NT}\right)^{-1}\frac{Z'X}{NT}\right)^{-1} \times\left(\frac{X'Z}{NT}\left(\frac{Z'\Omega Z}{NT}\right)^{-1}\frac{Z'Y}{NT}\right).$$

In this case, Y is P_{t+k}, X is a constant and ${}_tP_{i,t+k}$, and Z is $Z_{i,t}$. (The $Z_{i,t}$'s are instruments chosen from information available to forecaster i at time t.) Here the Ω matrix will have off-diagonal elements because of aggregate shocks and MA errors (for OLS, $\Omega=I$). To consistently estimate the covariance matrix we need to consistently estimate $(Z'\Omega Z/NT)$. If the errors in this model were i.n.i.d., then $(1/NT)\times$ $\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}Z_{i,t}'\hat{\varepsilon}_{t,k}^{i}\hat{\varepsilon}_{t,k}^{i}{}'Z_{i,t}$ would converge to a fixed matrix M, which is a consistent estimator of $(Z'\Omega Z/NT)$. If we merely had moving average errors, with no correlation of the errors across people, then with an MA(k) error, $(1/NT)\Sigma_{j=-k}^{k}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}$ $Z_{i,t}'\hat{\varepsilon}_{t,k}^{i}\hat{\varepsilon}_{t+j,k}^{i}{}'Z_{i,t+j}$ would be a consistent estimator. Note that as we increase the order of the MA error, the number of terms that we compute increases very slowly, so computation of this matrix is feasible.[14]

However, if errors are correlated across people, $(2k+1)N^2$ terms must be estimated to compute a general covariance matrix based on the orthogonality conditions from (2). Unfortunately, this calculation would exhaust the number of degrees of freedom in the data. However, we can construct a covariance matrix estimator that is consistent even in the presence of aggregate shocks if we make the following assumptions:

$$(8)\quad E\left(\varepsilon_{t+l,k}^{i}\varepsilon_{t+m,k}^{i}\right)=\begin{cases}\sigma_{|l-m|}, & \forall\, i,t,l,m,\ \text{s.t. } |l-m|\le k;\\ 0, & \text{otherwise.}\end{cases}$$

$$(9)\quad E\left(\varepsilon_{t+l,k}^{i}\varepsilon_{t+m,k}^{j}\right)=\begin{cases}\delta_{|l-m|}, & \forall\, i,j,t,l,m,\ \text{s.t. } |l-m|\le k;\\ 0, & \text{otherwise.}\end{cases}$$

This amounts to assuming that the data are not conditionally heteroskedastic and that no forecaster is systematically better than any other (i.e., σ_i's are the same for each individual).[15]

[12] We ignore problems with unit roots in this paper. Results of West (1988) show that in time-series like the deflator, in which trend terms dominate, standard asymptotic results for GMM estimators are correct.

[13] Zarnowitz (1985), arguing that pooled time-series cross-section regressions are misspecified, has not recognized that such regressions are $\sqrt{T}$ consistent. He does, however, recognize that OLS standard errors are inconsistent for these models.

[14] Newey and West (1987) discuss the limits on the growth of the number of terms that can be estimated.

[15] These assumptions are somewhat restrictive, but they seem reasonable. Although conditional heteroskedasticity exists in inflation, it is unlikely to exist with price-level forecasts. The findings of McNees (1975) and Zarnowitz (1967) that the accuracy of forecasters does not differ systematically supports assumption (9). Neither assumption can be formally tested because of the combination of aggregate shocks and moving average errors.

Given (8) and (9), we need only estimate the $(k+1)$ σ's and $(k+1)$ δ's, a feasible operation. Under these assumptions, our estimate of Ω will have the form

$$(10)\qquad \Omega=\begin{pmatrix} Q & R & \dots & R \\ R & Q & \dots & R \\ \vdots & \vdots & \ddots & \vdots \\ R & R & \dots & Q \end{pmatrix},$$

where

$$(11)\quad Q = \begin{pmatrix} \sigma_0 & \sigma_1 & \dots & \sigma_k & 0 & 0 & \dots & 0 & 0 \\ \sigma_1 & \sigma_0 & \dots & \sigma_{k-1} & \sigma_k & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_k & \sigma_{k-1} & \dots & \sigma_0 & \sigma_1 & \sigma_2 & \dots & 0 & 0 \\ 0 & \sigma_k & \dots & \sigma_1 & \sigma_0 & \sigma_1 & \dots & 0 & 0 \\ 0 & 0 & \dots & \sigma_2 & \sigma_1 & \sigma_0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \sigma_0 & \sigma_1 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \sigma_1 & \sigma_0 \end{pmatrix}$$

and

$$(12)\quad R = \begin{pmatrix} \delta_0 & \delta_1 & \dots & \delta_k & 0 & 0 & \dots & 0 & 0 \\ \delta_1 & \delta_0 & \dots & \delta_{k-1} & \delta_k & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \delta_k & \delta_{k-1} & \dots & \delta_0 & \delta_1 & \delta_2 & \dots & 0 & 0 \\ 0 & \delta_k & \dots & \delta_1 & \delta_0 & \delta_1 & \dots & 0 & 0 \\ 0 & 0 & \dots & \delta_2 & \delta_1 & \delta_0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \delta_0 & \delta_1 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \delta_1 & \delta_0 \end{pmatrix}$$

Given this estimate of Ω, we can compute $\hat{\beta}_{GMM}$ and its covariance matrix and test the restrictions implied by rationality.[16] No previous study has noted this form for $\hat{\beta}_{GMM}$. In fact, the limited panel-data work testing the rationality of forecasts has completely ignored this point.[17]

IV. Data

This study uses both panel survey data and aggregate time-series data. The survey data come from the ASA-NBER survey of economic forecasters who are members of the ASA Business and Economics Section. The panel is comprised of professional forecasters rather than research economists whose livelihood does not depend upon the accuracy of their forecasts.[18] Since they report to the survey the same forecasts that they sell on the market, these respondents have an economic incentive to make accurate predictions and we can assume that their survey responses are reasonably accurate measures of their expectations. Forecasters receive their surveys near the end of the first month of each quarter. They report their perceptions of the previous quarter's GNP deflator and their predictions for the current quarter and the following four quarters.[19] They respond by the end of the second month of the quarter.[20]

To test rationality, we compare individual quarterly forecasts with the deflator announcements released 45 days after the end of a given quarter. The deflator data used in

[16] Note that we must assume that $\varepsilon^i_{t-l,k}$ and $\varepsilon^j_{t+m,k}$ come from the same distribution. Thus we have symmetry rather than skew-symmetry in the R matrix.

[17] Since the dependent variable is the same for each forecaster, we cannot estimate a fixed time-effects model in this case. However, if there were cross-sectional variation in the dependent variable (as is true in the panel-data consumption literature), a fixed time-effects model could be estimated using the i.n.i.d. covariance estimator. With such a short panel, this is probably the preferred approach. For further discussion of this issue, see Runkle (1989).

[18] In fact, fewer than 5 percent of the respondents to the December 1970 survey were academics (Zarnowitz, 1974).

[19] They also have to regularly forecast all the major components of GNP in order to be included in the survey.

[20] It is not meaningless to ask about the deflator from the previous quarter since the 45-day NIPA release is made *after* the survey is mailed out. In fact, many respondents have already completed their forms when those numbers are released; we assume that those data are not in the forecasters' information set.

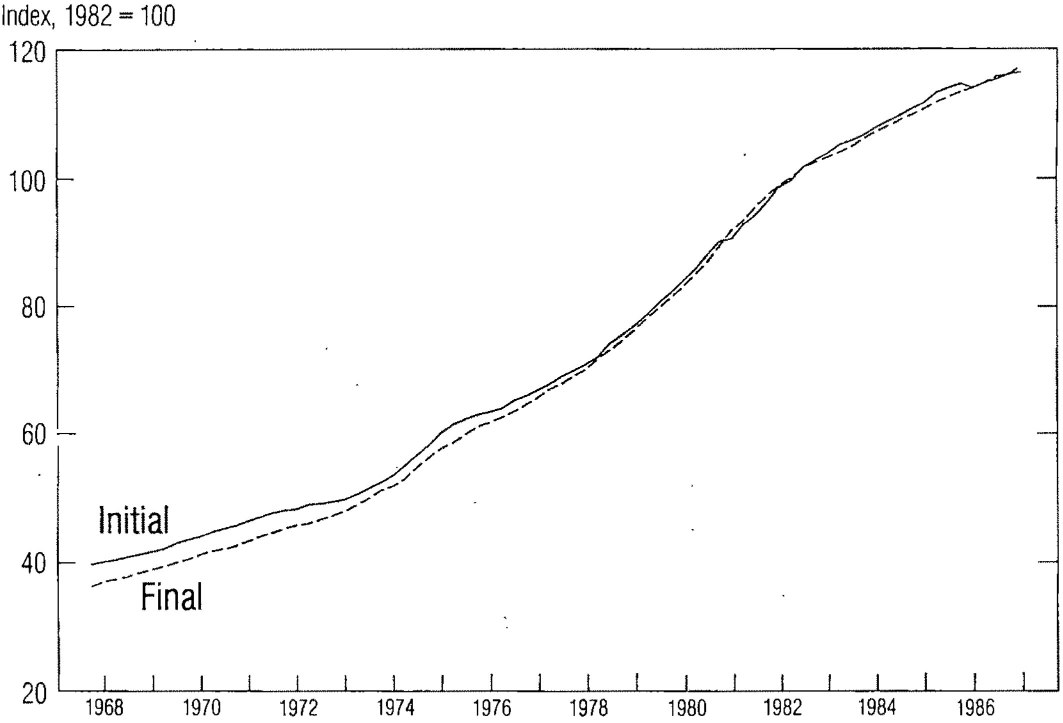


FIGURE 1. INITIAL AND FINAL GNP DEFLATOR DATA, 1968–1986

this study are often systematically revised. Forecasters generally do not know in advance (i.e., at time t) the nature of systematic data revisions occurring after the date on which they make their forecasts but before the date on which the NIPA announcements are made (i.e., time $t+1$). Often, the ASA-NBER survey asks forecasters for predictions that are further in the future than the next annual benchmark revision of the GNP deflator series. In these cases, forecasts differ systematically from announced values solely because of data revisions. We deal with this problem by excluding from the data all forecasts with horizons that extend beyond the date of systematic data revisions. We include only those observations for which there were no benchmark revisions from the time the survey was taken until the 45-day NIPA announcement was made. To do otherwise would severely bias our results against the hypothesis of rationality.

We assume that forecasters knew the GNP deflator announcements made in previous quarters. We allow for the fact that P_t, the current quarter's GNP deflator, is not necessarily known when ${}_tP_{t+1}$ is announced by assuming only that forecasters know their own perception of P_t, which is ${}_tP_t$. Note that the longer the forecast horizon, the more data we would lose because of our data screens. For this reason, we concentrate on one-step-ahead deflator forecasts and examine the rationality of price-level forecasts rather than inflation forecasts.

Many previous studies may be biased against the hypothesis of rationality because they assume that the final revisions of all variables contained in agents' information sets were known when forecasts were made. Figure 1 shows the effect of using revised rather than initial data. It shows the divergence between the 45-day announcement of the GNP deflator and the final data for the GNP deflator (as of September 1987), rescaled to reflect the benchmark revisions of 1976 and 1986. This figure shows that there are large systematic differences between these two price series.[21] Not only the level, but also the rate of growth of the series is affected.

Although we compare the forecasters' predictions to initial GNP deflator data in our study, it is important to note that using revised data in tests of unbiasedness is not necessarily inappropriate. Whether revised data should be used depends on two issues: First, do forecasters try to predict the initial or the revised data? Second, are there significant and predictable data revisions?

Obviously, if forecasters try to predict the revised data, those data should be the dependent variable in tests of unbiasedness. However, in the forecasts studied by Zarnowitz (1967) and McNees (1986), predictions were, on average, closer to the initial announcement than to the revision. This suggests that forecasters are, on average, trying to predict the initial announcement.

Regardless of which data the forecasters try to predict, data revisions should have little effect on tests of rationality unless they are significant and systematic. Since initial announcements of many variables, such as real GNP, are rational forecasts of the final data for those variables, the choice of dependent variable probably would have little effect on tests of the rationality of predictions of those variables.[22] But we show in Section V, Part C that the choice of data is very important in testing unbiasedness of forecasts of the GNP deflator, as one would expect from the evidence in Figure 1.

Although there are some circumstances in which it is permissible to use revised data for tests of unbiasedness, it is never permissible to use data unavailable to the forecasters as an independent variable in tests of efficiency. Thus, for example, if we wish to test whether forecasters properly adjust for their own past error $(P_{t-1} - {}_{t-2}P_{i,t-1})$, we should use the latest available revision of P_{t-1} before the forecaster made his prediction. We do this in our study. If we used final data for this test, the test would necessarily be incorrect because the final revision of P_{t-1} was not in the forecaster's information set when he made his prediction.

We use survey data from the fourth quarter of 1968 through the third quarter of 1986. We exclude periods containing the annual benchmark revisions, which usually occur in July, and forecasters who did not respond at least twenty times. Our sample contains 1613 observations. Some of our regressions use fewer observations because data are missing for other variables.

Like the survey data, our time-series data are reported as quarterly averages. The GNP deflator data come from the *Survey of Current Business*. Our M1 data come from the database of the Federal Reserve Board of Governors. Our nominal oil price statistics come from the Commerce Department's *Foreign Trade Statistics*.

V. Empirical Results

We address three questions in our empirical investigation. First, what is the covariance structure of the errors? Second, are the forecasters' predictions unbiased? Third, are the forecasters' predictions efficient?

All of the results presented are tests of rationality at the one-step-ahead horizon. Thus, our equation is

$$(13)\quad P_{t+1} = \alpha_0 + \alpha_1 {}_tP_{i,t+1} + \alpha_2 X_{i,t} + \varepsilon^i_{t,1},$$
$$E(\varepsilon^i_{t,1} \mid I_{i,t}) = 0.$$

To test for unbiasedness, we estimate regressions without $X_{i,t}$. Our null hypothesis in that case is that $\alpha_0 = 0$ and $\alpha_1 = 1$. To test for efficiency, we include some additional variable $X_{i,t}$ in the regressions. Our null hypothesis in that case is that $\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 0$.

A. *The Covariance Structure of Forecast Errors*

Two important questions arise about the covariance structure of the forecast errors.

[21] Note that Zarnowitz (1985) recognized that price-level data could not be compared across major benchmark revisions. No other author has recognized this. We discuss problems with Zarnowitz's data in Section V, Part B.

[22] See Mankiw and Shapiro (1986).

TABLE 1—TESTS FOR UNBIASEDNESS AND SERIAL INDEPENDENCE

Method	α_0	α_1	α_2	χ^2 for H_0	MA	Regressors	N
(1)	5.853	0.9967	–	21.66	0	$1, {}_tP_{i,t+1}$	1613
	(1.319)	(0.0008)	–	(0.00002)			
(2)	5.853	0.9967	–	1.226	0	$1, {}_tP_{i,t+1}$	1613
	(5.321)	(0.0031)	–	(0.5417)			
(3)	5.853	0.9967	–	0.900	1	$1, {}_tP_{i,t+1}$	1613
	(6.227)	(0.0036)	–	(0.6377)			
(4)	11.38	0.9943	0.3031	19.81	1	$1, {}_tP_{i,t+1}$	803
	(5.725)	(0.0032)	(0.0802)	(0.0002)		$P_t - {}_{t-1}P_{i,t}$	
(5)	7.078	0.9959	−0.0021	0.875	1	$1, {}_tP_{i,t+1}$	728
	(7.692)	(0.0044)	(0.1106)	(0.8315)		$P_{t-1} - {}_{t-2}P_{i,t-1}$	
(6)	7.756	0.9956	−0.0055	1.341	1	$1, {}_tP_{i,t+1}$	1111
	(6.720)	(0.0039)	(0.0695)	(0.7194)		${}_tP_{i,t} - {}_{t-1}P_{i,t}$	

Note: Standard errors are in parentheses under coefficients; significance levels are given under χ^2-statistics. H_0: $(\alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = 0)$.

First, are aggregate shocks important? Second, are forecast errors serially correlated? We address both issues in Table 1. Row 1 of Table 1 reports the results of a simple OLS regression to test the rationality of one-step-ahead forecast errors.[23] The results indicate that the hypothesis of unbiasedness is strongly rejected. However, that rejection is misleading because of the magnitude of aggregate shocks. In this OLS regression, the average covariance of the forecast errors made in a given period by two different forecasters is 58 percent of the variance of the average forecast error. This large covariance implies that a large percentage of forecast error variance comes from aggregate shocks, and thus we cannot assume that these errors are independent across forecasters. Row 2 reveals the importance of aggregate shocks. If we correct the standard errors in the OLS regression to reflect the cross-forecaster error covariance, our test results change dramatically. We do not reject the hypothesis of unbiasedness using a GMM estimator with a covariance matrix that accounts for aggregate shocks. In fact, the estimated standard error for α_1 increases from 0.0008 to 0.0031 when we do so. (Note that $\hat{\beta}_{GMM} = \hat{\beta}_{OLS}$ in this case, because the model is exactly identified, and the instruments are also the regressors.) All subsequent results account for the cross-sectional correlation of forecast errors that is created by aggregate shocks.

For results presented in Rows 1 and 2, we assume that forecast errors are not serially correlated. If forecast errors are not serially correlated, then the regression coefficient on the lagged forecast errors should not be significant. In Row 4, we report the results of including the lagged error as a regressor. The χ^2-statistic for the null hypothesis $(\alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = 0)$ clearly shows that that lagged error is correlated with the current forecast error. In fact, there is a very strong MA(1) correlation (the MA parameter is 0.3031).

Rejecting serial independence might cause us to question the rationality of the price-level forecasts.[24] However, we need to conduct additional tests to see whether this serial correlation implies that we could improve the forecasters' predictions using data they knew when they made their forecasts. We test whether two such pieces of data could improve our forecasts. First, as shown in Row 5, we test whether the second lagged error is significant in predicting future

[23] This test is similar to Figlewski and Wachtel's (1981).

[24] Zarnowitz (1985) draws this inference.

prices. The value for the χ^2-statistic is extremely low, indicating that this error has no power to predict future prices. This supports the hypothesis of rationality, since the second lagged errors were definitely known by forecasters when they made their predictions, ${}_tP_{i,t+1}$. Second, as reported in Row 6, we test whether the *perceived* lagged forecast error helps predict future prices. Since the survey elicits not only the predictions ${}_tP_{i,t+1}$ but also ${}_tP_{i,t}$ (the forecasters' perceptions of P_t), we can construct the forecasters' perceptions at time t of the one-step-ahead forecast errors they made from time $t-1$ to time t. Since these perceived errors are certainly in the forecasters' information sets, they should not help to predict future prices. The χ^2-statistic in Row 6 shows that the perceived forecast errors do not improve price forecasts. Thus, only the unperceived part of lagged forecast errors helps to predict future prices. (Remember that P_t is not known when the forecast ${}_tP_{i,t+1}$ is made.) The presence of MA(1) errors arising from unperceived lagged errors is compatible with the hypothesis of rationality. All subsequent regressions assume an MA(1) error structure.

B. *Unbiasedness Tests*

In Row 3, we present our OLS results assuming MA(1) errors and correcting standard errors for the effect of aggregate shocks. The estimate of α_0 is 5.853 (with a standard error of 6.227), and the estimate of α_1 is 0.9967 (with a standard error of 0.0036).[25] The value of the χ^2 test for unbiasedness is 0.900. Since this test statistic has a probability value of 0.6377, we cannot reject the null hypothesis of rationality.

Such strong support for rationality seems incompatible with Zarnowitz's (1985) rejection of unbiasedness using the same data. The only obvious differences between Zarnowitz's study and ours are that he looks at the unbiasedness of inflation forecasts and does not pool the data. But there is another difference that explains his rejection of unbiasedness: Zarnowitz uses revised data. He uses as his actual inflation series the final available data before the benchmark revisions of January 1976 and December 1980. We find that the annual July data revisions of the GNP deflator series by the Department of Commerce appear to have large systematic components. Therefore we use as *actual* data the final values available *before* each annual July revision.

Using revised data, Zarnowitz finds the striking result that, for one-quarter-ahead inflation forecasts, the hypothesis of rational expectations can be rejected for 27 percent of the forecasters at the 5 percent level and for 69 percent of the forecasters at the 10 percent level. Over the 1968 to 1979 period, he finds a mean error of -0.16 percent for one-quarter-ahead inflation forecasts. Unfortunately, these results arise from Zarnowitz's use of revised rather than initial data.[26]

We reconstructed the data that Zarnowitz used for P_{t+1} from different issues of *Business Conditions Digest* and adjusted the sample so that all our regressions would include the same observations. Using Zarnowitz's approach, we then estimated individual regressions for each forecaster in our sample using both Zarnowitz's data and our unrevised data. The difference in the results is striking.[27] The following table summarizes our findings.

Number (and Percentage) of Individual Forecasters for Whom Unbiasedness Is Rejected

Data Set	5 Percent Level	1 Percent Level
Zarnowitz	27 (45.0 percent)	13 (21.7 percent)
Initial Data	8 (13.3 percent)	1 (1.6 percent)

This table shows that using initial data as the dependent variable results in test statistics that are far more favorable to the hy-

[25] The survey data are reported in 1000's, which is why our estimated constant is so large.

[26] Zarnowitz recognized that the use of revised data might affect his inference, but he believed the effect would not be large.

[27] We present a complete set of these individual regressions in the Appendix.

pothesis of unbiasedness. We believe that these data are a more accurate reflection of what the forecasters were trying to predict, and of what their information sets contained, than are the revised data used by Zarnowitz.

Our use of levels instead of inflation does not account for the difference in the results because rejections at the 5 percent level increase from 27 percent to 45 percent of the sample when we use the price level rather than the inflation rate in our regressions. Clearly our use of initial data accounts for the difference of our results from Zarnowitz's and our failure to reject rationality.

Testing whether the initial announcements of the deflator are rational forecasts of the revised data that Zarnowitz used further indicates the difference between our price data and Zarnowitz's. If there were no systematic revisions in the data, the 45-day GNP deflator announcement should be a rational forecast of the revised data. In that case, if we regress Zarnowitz's data on a constant and the 45-day announcement, the coefficient on the constant should be zero and the coefficient on the announcement should be one.[28] But when we computed that regression, the test statistic for the hypothesis that those coefficients had their assumed values was 15.92. Since this test statistic should be distributed as a χ_2^2 random variable under the null hypothesis, we must reject the notion that the 45-day deflator announcements are rational forecasts of the data used by Zarnowitz.[29]

As we mentioned before, using revised data may not have much of an effect on tests of unbiasedness if the initial announcement was a rational forecast of the final data, as was probably true with Zarnowitz's other data series. But for the GNP deflator, the choice of dependent variable is important. We cannot determine a priori whether forecasters are trying to forecast the initial or the revised data, but the fact that their forecasts are unbiased predictors of the initial data and biased predictors of the revised data suggests that they are trying to predict the initial announcement. Thus, it seems more appropriate to use the 45-day announcement as the dependent variable.

To the best of our knowledge, all of the other authors in the inflation-forecast literature have used revised data as the independent variable for their tests and as a conditioning set in their tests of rationality.[30] Since the revised data are not in the forecasters' information sets when they make their forecasts, other researchers' tests of rationality are biased toward rejection.

C. *Efficiency Tests*

Since we cannot reject unbiasedness, we must test the further implication of rational expectations that forecasts be efficient, that is, that no readily available information could have improved forecast accuracy. This involves testing the hypothesis that $\alpha_0 = 0$, $\alpha_1 = 1$, and $\alpha_2 = 0$ in equation (13). Tables 2, 3, and 4 give the results of our efficiency tests.

First note that none of the regressions reported in Table 2 rejects the coefficient restrictions implied by efficiency (H_0). In Rows 1 through 5, respectively, the following variables are shown not to improve price forecasts: the forecaster's perception of the price level at the time of his forecast (${}_tP_{i,t}$), the lagged price level (P_{t-1}), the forecaster's previous one-step-ahead prediction (${}_{t-1}P_{i,t}$), the lagged value of the forecaster's error in perceiving the price level ($P_{t-1} - {}_{t-1}P_{i,t-1}$), and the forecaster's perception of another forecaster's lagged error (${}_tP_{i,t} - {}_{t-1}P_{j,t}$). Remarkably, none of these implications of rationality is rejected.

[28] Mankiw, Runkle, and Shapiro (1984) devised this test.

[29] We also adopted the suggestion of an anonymous referee to test whether this result holds true in each half of the data. It does. We reject the hypothesis that the 45-day announcement is a rational forecast of Zarnowitz's data at the 5 percent level in each half-sample.

[30] Zarnowitz (1967) compares the annual forecasts of GNP and several other variables to their initial announcements for the years 1953–1963. He recognizes the importance of studying price forecasts (p. 138), but does not do so in that study.

TABLE 2—TESTS OF RESTRICTIONS IMPLIED BY EFFICIENCY

Method	α_0	α_1	α_2	χ^2 for H_0	MA	Regressors	N
(1)	4.542	0.8097	0.1906	3.847	1	1, ${}_tP_{i,t+1}$,	1599
	(6.929)	(0.1012)	(0.1024)	(0.2785)		${}_tP_{i,t}$	
(2)	7.243	0.8797	0.1200	4.012	1	1, ${}_tP_{i,t+1}$,	1079
	(7.646)	(0.0645)	(0.0659)	(0.2602)		P_{t-1}	
(3)	7.471	0.9112	0.0859	2.331	1	1, ${}_tP_{i,t+1}$,	1119
	(7.132)	(0.0676)	(0.0678)	(0.5066)		${}_{t-1}P_{i,t}$	
(4)	7.876	0.9956	0.0200	1.259	1	1, ${}_tP_{i,t+1}$,	797
	(7.034)	(0.0041)	(0.2546)	(0.7390)		$P_{t-1} - {}_{t-1}P_{i,t-1}$	
(5)	8.343	0.9953	-0.0823	2.478	1	1, ${}_tP_{i,t+1}$,	1457
	(6.985)	(0.0040)	(0.0701)	(0.479)		${}_tP_{i,t} - {}_{t-1}P_{j,t}$	

Note: Standard errors are in parentheses under coefficients; significance levels are given under χ^2-statistics. H_0: $(\alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = 0)$.

TABLE 3—FURTHER TESTS OF EFFICIENCY

Method	α_0	α_1	α_2	χ^2 for H_0	MA	Regressors	N
(1)	3.092	0.9992	-0.0009	1.0567	1	1, ${}_tP_{i,t+1}$,	1613
	(7.549)	(0.0056)	(0.0017)	(0.7875)		$P_{O,t-1}$	
(2)	0.7887	1.0012	-0.0016	1.6601	1	1, ${}_tP_{i,t+1}$,	1569
	(7.523)	(0.0055)	(0.0016)	(0.6458)		$P_{O,t-2}$	
(3)	4.481	1.0003	-0.0014	1.964	1	1, ${}_tP_{i,t+1}$,	1613
	(6.011)	(0.0045)	(0.0012)	(0.5799)		$M1_{t-1}$	
(4)	4.704	1.0003	-0.0014	1.868	1	1, ${}_tP_{i,t+1}$,	1569
	(6.214)	(0.0047)	(0.0013)	(0.6002)		$M1_{t-2}$	

Note: Standard errors are in parentheses under coefficients; significance levels are given under χ^2-statistics. H_0: $(\alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = 0)$.

TABLE 4—JOINT TEST OF EFFICIENCY

$P_{t+1} = \alpha_0 + \alpha_{1\,t}P_{i,t+1} + \alpha_2 P_{t-1} + \alpha_3(P_{t-1} - {}_{t-2}P_{i,t-1}) + \alpha_4({}_tP_{i,t} - {}_{t-1}P_{i,t}) + \alpha_5 P_{O,t-1} + \alpha_6 M1_{t-1} + \varepsilon_{i,t+1}$	
α_0	21.882
	(11.451)
α_1	0.806
	(0.073)
α_2	0.191
	(0.073)
α_3	0.022
	(0.011)
α_4	0.005
	(0.082)
α_5	0.006
	(0.003)
α_6	-0.005
	(0.004)
No. of MA Terms	1
χ^2-Statistic	8.288
Significance Level	0.308
N	576

Note: Standard errors are in parentheses.

Two variables that varied greatly, and presumably had a large effect on the price level during the sample period, are the money supply and the price of oil. Economic forecasters were accused of systematically underestimating the effects of these variables on price behavior during the 1970s and 1980s. Table 3 examines whether forecasters fully adjusted their predictions to changes in the money supply and oil prices. This table presents regressions that include the previous two lagged values of nominal crude oil prices $(P_{O,t-1}, P_{O,t-2})$ and M1 $(M1_{t-1}, M1_{t-2})$ as dependent variables. The coefficient restrictions implied by rationality are not rejected in any of these regressions.

Although the results in Tables 1, 2, and 3 show strong support for forecast rationality, a joint test of forecast rationality using the most important variables from those tables

TABLE 5—TESTS OF FALSE EFFICIENCY RESTRICTIONS

Method	α_0	α_1	α_2	χ^2 for H_0	MA	Regressors	N
(1)	8.602	0.7093	0.2908	10.76	1	$1, {}_tP_{i,t+1}$,	1129
	(7.272)	(0.0970)	(0.0978)	(0.0131)		P_t.	
(2)	0.6950	0.0567	0.9431	2353.7	1	$1, {}_tP_{i,t+1}$,	1172
	(1.380)	(0.0195)	(0.0196)	(0.0000)		${}_{t+1}P_{i,t+1}$	
(3)	4.268	0.5457	0.4457	110.2	1	$1, {}_tP_{i,t+1}$,	1178
	(4.349)	(0.0441)	(0.0438)	(0.0000)		${}_{t+1}P_{i,t+2}$	
(4)	10.12	0.9949	0.6891	19.64	1	$1, {}_tP_{i,t+1}$,	1120
	(5.760)	(0.0033)	(0.1719)	(0.0002)		$P_t - {}_tP_{i,t}$	

Note: Standard errors are in parentheses under coefficients; significance levels are given under χ^2-statistics. H_0: $(\alpha_0 = 0, \alpha_1 = 1, \alpha_2 = 0)$.

would provide even more convincing evidence that these forecasts are rational. Table 4 gives the results of a test of rationality in which the P_{t+1} was regressed on a constant, the one-step-ahead forecast ${}_tP_{i,t+1}$, the lagged price level P_{t-1}, the lagged perceived forecast error ${}_tP_{i,t} - {}_{t-1}P_{i,t}$, the second lagged forecast error $P_{t-1} - {}_{t-2}P_{i,t-1}$, the lagged nominal crude oil price $P_{O,t-1}$, and the lagged money supply $M1_{t-1}$. The χ^2 test statistic for this regression shows that the hypothesis of efficiency cannot be rejected.[31]

Given the strong support that the results in Tables 1–4 have provided for the hypothesis of forecast rationality, we might wonder whether these tests have any power. Table 5 addresses this issue. It displays the results of regressions that include variables that were not in a forecaster's information set when he made his price forecasts. In fact, the χ^2-statistics show that we can reject the test for forecast rationality for each of these variables. Row 1 of Table 5 shows regression results including the contemporaneous (but unknown) price level $(P_{i,t})$ as a regressor. Row 2 shows the results of including the forecaster's perception of the price level at the end of the period $({}_{t+1}P_{t+1})$. Row 3 results include the forecaster's next one-step-ahead prediction $({}_{t+1}P_{i,t+2})$. Results in Row 4 include the forecaster's contemporaneous error in perceiving the price level $(P_t - {}_tP_{i,t})$. The test of efficiency is rejected in every case. Thus, these tests have sufficient power to reject invalid restrictions.

The results in Row 4 are particularly instructive. The forecaster's perceived forecast error from the previous period $({}_tP_{i,t} - {}_{t-1}P_{i,t})$ is orthogonal to the current forecast error (see Table 1, Row 6. However, since his unperceived forecast error $(P_t - {}_tP_{i,t})$ is not orthogonal to his forecast error (see Table 5, Row 4), neither is the actual lagged forecast error, because it is the sum of the perceived and unperceived components (see Table 1, Row 4).[32]

The results in Tables 1–5 strongly confirm the hypothesis that survey forecasts are rational. The variables expected to improve price forecasts do so, and the variables not expected to improve price forecasts do not. The coefficient restrictions rationality implies cannot be rejected when the standard errors are correctly computed.

The rational expectations hypothesis implies that only differential private informa-

[31] Although the money supply and oil prices certainly do not exhaust the list of information available to the forecasters, the fact that forecasters seem to have efficiently used this information suggests that the forecasts are rational. Mullineaux (1980) found similar results in testing whether the mean Livingston forecaster efficiently incorporated data about money growth.

[32] Boschen and Grossman (1982) first made this distinction between perceived and unperceived forecast errors.

tion should explain differences among individual forecasts. Our final set of tests checks whether differences among the forecasts of different forecasters can be explained by publicly available information. We tested whether the lagged growth rate of M1 or the lagged growth rate of oil prices could explain the difference between any given forecaster's prediction error and the mean prediction error for that period. If either of those variables is correlated with idiosyncratic error components for some forecasters, then those forecasters did not fully account for that publicly available information in making their forecasts.[33] In only one of the 60 individual regressions is the rate of money growth significant at the 5 percent level in explaining the forecaster's idiosyncratic errors. In only 4 of the 60 individual regressions is the rate of oil price growth significant at the 5 percent level in explaining those errors. This suggests that heterogeneity in individual forecasts is not due to differential accuracy among forecasters in predicting the effects of money growth or oil price growth on future prices.

VI. Conclusion

Our results indicate that survey respondents' forecasts of the GNP deflator are both unbiased and efficient—and therefore rational. We demonstrate the importance of accounting for aggregate shocks in order to conduct correct statistical inference in panel data models. We find that failure to account for such shocks when estimating the covariance matrix for the regression estimates leads to false rejection of the hypothesis of rationality. We develop a covariance matrix consistent in the presence of aggregate shocks that leads to strong affirmation of rationality. We also demonstrate that although a strong moving average component is present in agents' one-step-ahead forecast error, it does not imply irrationality, because it arises from imperfect information regarding the current price level. Interestingly, this implies that the class of rational expectations models in which only unanticipated inflation can cause output to deviate from trend can explain persistence in output. Also, we show that using revised rather than initial price data to test the rationality of forecasts can greatly affect the results of those tests. In fact, Zarnowitz, who uses the same forecast data we do, rejects rationality for inflation forecasts solely because he uses revised data. Finally, we find that we cannot reject the strong implication of rational expectations that forecast differences should only result from asymmetric information. Readily available data do not explain the differences among individual forecasts.

Yet our results are not conclusive evidence for rational expectations. We find that expert forecasters are rational, but other people may not be. Still, our results and methodological discussion cast grave doubts upon the assumptions, methods, and conclusions in the existing literature. And, although the support we provide for the rational expectations hypothesis is limited, it takes on added importance when viewed from the perspective that almost the entire existing literature has rejected the rationality of price forecasts, even from professional forecasters. Hence, our results can be viewed as salvaging the possibility that the rational expectations hypothesis is empirically valid, and reopening the debate on this subject.

Finally, our finding of the importance of aggregate shocks should prompt other economists who explore macroeconomic questions with panel data to reexamine their statistical assumptions. It seems unlikely that the residuals in micro consumption and labor studies are independent within each cross section. Failure to adjust for this dependence could lead to incorrect statistical inference.[34]

[33] The idiosyncratic component is $\varepsilon^i_{t,1} - \bar{\varepsilon}_{t,1}$.

[34] Altug and Miller (1990) find aggregate shocks very important in their study of consumption and labor supply.

APPENDIX

Individual Test Results Using Zarnowitz's Data

NBER I.D.	α_0	α_1	H_0	NBER I.D.	α_0	α_1	H_0
3	2.4898	1.0011	2.8829	7	−8.2951	1.0087	11.229
	(14.961)	(0.00974)	(0.23658)		(12.343)	(0.00796)	(0.00364)
8	4.4746	0.99956	1.8702	14	8.3869	0.99883	19.312
	(22.851)	(0.01478)	(0.39255)		(10.713)	(0.00709)	(0.00006)
15	11.962	0.99445	5.4328	21	−1.1597	1.0056	4.0226
	(7.9520)	(0.00496)	(0.06611)		(21.361)	(0.01251)	(0.01338)
22	−3.9781	1.0052	3.6053	23	20.768	0.99066	3.0006
	(18.836)	(0.01255)	(0.16486)		(14.039)	(0.00757)	(0.22307)
27	24.443	0.98641	5.6445	30	19.143	0.99200	1.3584
	(10.298)	(0.00590)	(0.05947)		(20.582)	(0.01011)	(0.50703)
31	20.897	0.99162	12.272	32	9.2309	0.9970	9.4499
	(11.312)	(0.00670)	(0.00216)		(13.606)	(0.00890)	(0.00887)
34	−0.72462	1.0037	7.5493	38	15.030	0.99281	7.2617
	(14.304)	(0.00909)	(0.02295)		(13.080)	(0.00853)	(0.02649)
39	14.890	0.99332	4.9309	40	18.354	0.99225	4.1992
	(11.189)	(0.00688)	(0.08497)		(13.503)	(0.00745)	(0.12251)
42	1.7853	1.0026	9.9783	43	30.891	0.98747	20.689
	(11.301)	(0.00716)	(0.00681)		(11.097)	(0.00605)	(0.00003)
44	−6.3401	1.0059	2.0942	46	8.4090	0.99806	5.4088
	(16.299)	(0.00979)	(0.35096)		(16.835)	(0.01120)	(0.06691)
47	2.5519	0.99761	0.23846	48	10.198	0.99800	12.409
	(20.115)	(0.01140)	(0.88760)		(16.415)	(0.01102)	(0.00202)
49	−6.8452	1.0079	28.547	51	26.053	0.98454	11.952
	(8.0293)	(0.00522)	(0.0000006)		(7.5482)	(0.00452)	(0.00254)
54	−13.488	1.0125	18.141	57	10.081	0.99654	5.8440
	(11.293)	(0.00725)	(0.00011)		(15.853)	(0.01026)	(0.05383)
59	9.6887	0.99712	6.6950	60	17.905	0.98964	4.9027
	(13.193)	(0.00866)	(0.03517)		(8.8023)	(0.00477)	(0.08618)
61	15.960	0.99276	4.8359	62	9.8688	0.99707	2.4453
	(9.3530)	(0.00537)	(0.08911)		(14.246)	(0.00728)	(0.29445)
64	4.8450	0.99910	6.8349	65	11.691	0.99563	6.5628
	(10.514)	(0.00692)	(0.03280)		(9.0874)	(0.00518)	(0.03758)
66	13.732	0.99389	2.6122	67	22.335	0.98815	6.3599
	(10.254)	(0.00553)	(0.27087)		(8.8707)	(0.00490)	(0.04159)
68	29.936	0.98663	6.8446	69	14.398	0.99324	2.5578
	(13.878)	(0.00732)	(0.03264)		(11.620)	(0.00678)	(0.27834)
70	8.5655	0.99709	7.1223	72	9.0130	0.99754	4.4448
	(7.0671)	(0.00398)	(0.02841)		(11.390)	(0.00663)	(0.10835)
73	20.168	0.99007	13.373	75	23.365	0.98999	7.3882
	(12.249)	(0.00801)	(0.00125)		(12.945)	(0.00751)	(0.02487)
77	−15.027	1.0112	1.5137	78	2.2050	1.0015	6.4330
	(18.470)	(0.01204)	(0.46913)		(11.446)	(0.00724)	(0.04010)
82	18.430	0.99119	6.5981	84	11.607	0.99486	1.8743
	(8.9620)	(0.00517)	(0.03692)		(11.200)	(0.00598)	(0.39174)
86	7.0478	0.99869	7.0460	87	8.8439	0.99406	1.7829
	(9.9970)	(0.00598)	(0.02951)		(10.752)	(0.00598)	(0.41007)
89	15.341	0.99269	4.2050	93	27.050	0.98559	3.7746
	(11.190)	(0.00666)	(0.12215)		(15.571)	(0.00925)	(0.15148)
96	2.4145	0.99318	0.28821	98	11.779	0.99478	8.6409
	(15.160)	(0.00915)	(0.86580)		(7.5815)	(0.00461)	(0.01329)
102	14.569	0.99180	0.77244	108	−31.562	1.0250	12.634
	(22.167)	(0.01409)	(0.67962)		(22.838)	(0.01432)	(0.00181)
109	−21.115	1.0176	11.217	112	−24.921	1.0160	2.0212
	(20.057)	(0.01248)	(0.00367)		(19.960)	(0.01215)	(0.36401)
125	−5.9854	1.0002	3.3509	138	−17.728	1.0132	5.8484
	(26.976)	(0.01684)	(0.18722)		(13.384)	(0.00845)	(0.05371)
144	−16.815	1.0114	2.1684	145	−6.3399	1.0055	1.5882
	(14.334)	(0.00898)	(0.33816)		(22.023)	(0.01328)	(0.45199)

NBER I.D.	α_0	α_1	H_0	NBER I.D.	α_0	α_1	H_0
148	−35.524	1.0289	6.4511	158	−3.1309	1.0053	5.9090
	(41.462)	(0.02586)	(0.03973)		(20.486)	(0.01267)	(0.05210)

Note: Standard errors are in parentheses under coefficients; significance levels are given under χ^2-statistics for H_0.

Individual Test Results Using Initially Released Data

NBER I.D.	α_0	α_1	H_0	NBER I.D.	α_0	α_1	H_0
3	9.1382	0.99450	0.47508	7	0.48552	1.0005	0.91773
	(14.126)	(0.00920)	(0.78857)		(10.056)	(0.00648)	(0.63200)
8	6.7419	0.99550	0.12371	14	11.003	0.99423	5.8201
	(21.129)	(0.01302)	(0.94002)		(7.8468)	(0.00519)	(0.05447)
15	3.3783	0.99820	0.28922	21	−0.38588	1.0013	0.37485
	(7.4004)	(0.00462)	(0.86536)		(15.913)	(0.00932)	(0.82909)
22	−13.857	1.0102	1.4730	23	12.695	0.99288	2.0525
	(16.682)	(0.01111)	(0.4788)		(9.2274)	(0.00497)	(0.35835)
27	17.579	0.98852	8.4831	30	1.5236	0.99892	0.08735
	(7.7145)	(0.00442)	(0.01438)		(14.447)	(0.00710)	(0.95726)
31	19.276	0.99011	6.4269	32	−3.4191	1.0037	2.4887
	(8.9184)	(0.00528)	(0.04022)		(12.352)	(0.00808)	(0.28812)
34	4.6144	0.99801	1.1179	38	3.3336	0.99895	1.5414
	(11.326)	(0.00720)	(0.57182)		(11.519)	(0.00751)	(0.46268)
39	8.80857	0.99580	1.3134	40	12.607	0.99394	2.0482
	(8.7244)	(0.00536)	(0.51857)		(10.164)	(0.00561)	(0.35912)
42	−5.4555	1.0048	2.2748	43	19.672	0.99235	11.113
	(9.8857)	(0.00626)	(0.32066)		(10.432)	(0.00569)	(0.00386)
44	−11.053	1.0070	0.70276	46	−4.6930	1.0045	1.1088
	(14.223)	(0.00854)	(0.70371)		(14.882)	(0.00990)	(0.57441)
47	−3.2375	0.99910	2.0598	48	−0.20145	1.0031	5.7632
	(18.572)	(0.01052)	(0.35705)		(14.478)	(0.00972)	(0.05604)
49	−5.6613	1.0048	6.0616	51	14.644	0.99033	7.3626
	(6.0361)	(0.00392)	(0.04828)		(6.2969)	(0.00377)	(0.02519)
54	−1.1542	1.0018	2.2354	57	−5.4562	1.0048	1.5203
	(8.6148)	(0.00553)	(0.32704)		(13.974)	(0.00904)	(0.46760)
59	−1.8680	1.0027	1.7603	60	9.2780	0.99340	6.7461
	(11.181)	(0.00734)	(0.41471)		(6.7636)	(0.00367)	(0.03428)
61	7.5849	0.99623	0.96154	62	−3.8786	1.0023	0.28313
	(8.5865)	(0.00493)	(0.61831)		(9.3386)	(0.00477)	(0.86801)
64	2.2385	0.99869	0.08927	65	4.3618	0.99774	0.47456
	(9.5383)	(0.00628)	(0.95635)		(6.7932)	(0.00387)	(0.78877)
66	5.1299	0.99726	0.55304	67	11.565	0.99224	4.8101
	(6.9177)	(0.00373)	(0.75842)		(7.6704)	(0.00424)	(0.09026)
68	19.760	0.98953	3.9391	69	12.249	0.99186	3.3008
	(10.024)	(0.00529)	(0.13952)		(8.4623)	(0.00494)	(0.19200)
70	3.4932	0.99804	0.41833	72	5.3485	0.99709	0.43405
	(5.4228)	(0.00305)	(0.81126)		(8.3018)	(0.00483)	(0.80491)
73	7.1518	0.99709	5.6098	75	23.011	0.98691	7.2293
	(9.7200)	(0.00636)	(0.06051)		(8.5596)	(0.00496)	(0.02693)
77	−4.7178	1.0019	0.94257	78	2.0963	0.99917	0.26855
	(14.526)	(0.00947)	(0.62420)		(10.034)	(0.00635)	(0.87435)
82	11.719	0.99338	3.2729	84	2.6116	0.99839	0.18663
	(6.4776)	(0.00374)	(0.19467)		(8.3672)	(0.00447)	(0.91091)
86	10.445	0.99364	1.9437	87	1.8489	0.99634	5.6542
	(7.6236)	(0.00456)	(0.37837)		(11.063)	(0.00611)	(0.05918)
89	7.2372	0.99551	0.73126	93	17.439	0.99021	1.2036
	(8.9314)	(0.00532)	(0.69376)		(16.295)	(0.00968)	(0.54782)
96	6.0837	0.99435	5.5642	98	6.4943	0.99638	1.2734
	(9.1806)	(0.00554)	(0.06191)		(6.2992)	(0.00383)	(0.52903)

NBER I.D.	α_0	α_1	H_0	NBER I.D.	α_0	α_1	H_0
102	15.938	0.98915	0.79484	108	−20.179	1.0159	5.3174
	(21.261)	(0.01352)	(0.67205)		(21.714)	(0.01362)	(0.07004)
109	−15.376	1.0120	4.8020	112	−9.5870	1.0046	1.3252
	(17.518)	(0.01090)	(0.09063)		(17.037)	(0.01037)	(0.51552)
125	0.11363	0.99449	9.0306	138	−9.1007	1.0054	0.80104
	(25.318)	(0.01580)	(0.01094)		(11.147)	(0.00704)	(0.66997)
144	−8.9352	1.0045	2.0037	145	−2.8352	1.0020	0.05842
	(11.625)	(0.00728)	(0.36719)		(20.514)	(0.01237)	(0.97121)
148	−34.150	1.0261	3.9021	158	2.4112	0.99915	0.31752
	(39.627)	(0.02471)	(0.14212)		(17.144)	(0.01060)	(0.85320)

Note: Standard errors are in parentheses under coefficients; significance levels are given under χ^2-statistics for H_0.

REFERENCES

Altug, Sumru and Miller, Robert A., "Household Choices in Equilibrium," *Econometrica*, forthcoming.

Arrow, Kenneth J., "The Future and the Present in Economic Life," *Economic Inquiry*, April 1978, *16*, 157–69.

Boschen, John F. and Grossman, Herschel I., "Tests of Equilibrium Macroeconomics Using Contemporaneous Monetary Data," *Journal of Monetary Economics*, November 1982, *10*, 309–33.

Brown, Bryan W. and Maital, Shlomo, "What Do Economists Know? An Empirical Study of Experts' Expectations," *Econometrica*, March 1981, *49*, 491–504.

Carlson, John A., "A Study of Price Forecasts," *Annals of Economic and Social Measurement*, Winter 1977, *6*, 27–56.

Caskey, John, "Modeling the Formation of Price Expectations: A Bayesian Approach," *American Economic Review*, September 1985, *75*, 768–76.

Chamberlain, Gary, "Panel Data," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. II, Amsterdam: North-Holland, 1984, pp. 1247–1318.

Cukierman, Alex, "Measuring Inflationary Expectations: A Review Essay," *Journal of Monetary Economics*, March 1986, *17*, 315–24.

DeLeeuw, Frank and McKelvey, Michael J., "Price Expectations of Business Firms," *Brookings Papers on Economic Activity*, 1981, *1*, 299–313.

______ **and** ______, "Price Expectations of Business Firms: Bias in the Short and Long Run," *American Economic Review*, March 1984, *74*, 99–110.

Figlewski, Stephen and Wachtel, Paul, "The Formation of Inflationary Expectations," *Review of Economics and Statistics*, February 1981, *63*, 1–10.

______ **and** ______, "Rational Expectations, Informational Efficiency, and Tests Using Survey Data: A Reply," *Review of Economics and Statistics*, August 1983, *65*, 529–31.

Frankel, Jeffrey A. and Froot, Kenneth A., "Using Survey Data to Test Standard Propositions Regarding Exchange Rate Expectations," *American Economic Review*, March 1987, *77*, 133–53.

Friedman, Benjamin M., "Survey Evidence on the 'Rationality' of Interest Rate Expectations," *Journal of Monetary Economics*, October 1980, *6*, 453–65.

Gramlich, Edward M., "Models of Inflation Expectations Formation," *Journal of Money, Credit, and Banking*, May 1983, *15*, 155–73.

Hirsch, Albert A. and Lovell, Michael C., *Sales Anticipations and Inventory Behavior*, New York: Wiley & Sons, 1969.

Keane, Michael P. and Runkle, David E., "Are Economic Forecasts Rational?" *Federal Reserve Bank of Minneapolis Quarterly Review*, Spring 1989, *13*, 26–33.

Keynes, John Maynard, *The General Theory of Employment, Interest, and Money*. London: Macmillan, 1936.

Leonard, Jonathan S., "Wage Expectations in the Labor Market: Survey Evidence on Rationality," *Review of Economics and Statistics*, February 1982, *64*, 57–61.

Lovell, Michael C., "Tests of the Rational Expectations Hypothesis," *American Economic Review*, March 1986, *76*, 110–24.

Lucas, Robert E., Jr. and Rapping, Leonard A., "Real Wages, Employment, and Inflation," *Journal of Political Economy*, September/October 1969, *77*, 721–54.

Mankiw, N. Gregory, Runkle, David E. and Shapiro, Matthew D., "Are Preliminary Announcements of the Money Stock Rational Forecasts?" *Journal of Monetary Economics*, July 1984, *14*, 15–27.

______ **and Shapiro, Matthew D.**, "News or Noise: An Analysis of GNP Revisions," *Survey of Current Business*, May 1986, *66*, 20–25.

McNees, Stephen K., "An Evaluation of Economic Forecasts," *New England Economic Review* (Federal Reserve Bank of Boston), November/December 1975, 3–39.

______, "Estimating GNP: The Trade-Off Between Timeliness and Accuracy," *New England Economic Review* (Federal Reserve Bank of Boston), January/February 1986, 3–10.

Mullineaux, Donald J., "On Testing for Rationality: Another Look at the Livingston Price Expectations Data," *Journal of Political Economy*, April 1978, *86*, pt. 1, 329–36.

______, "Inflation Expectations and Money Growth in the United States," *American Economic Review*, March 1980, *70*, 149–61.

Muth, John F., "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

______, "Short Run Forecasts of Business Activity," paper presented at the joint Pittsburgh Meetings of the Eastern Economic Association–International Society for Inventory Research, March 1985.

Newey, Whitney K. and West, Kenneth D., "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, May 1987, *55*, 703–08.

Pearce, Douglas K., "Comparing Survey and Rational Measures of Inflation," *Journal of Money, Credit, and Banking*, November 1979, *11*, 447–56.

Pesando, James E., "A Note on the Rationality of the Livingston Price Expectations," *Journal of Political Economy*, August 1975, *83*, 849–58.

Pigou, Arthur C., *Industrial Fluctuations*, London: Macmillan, 1927.

Prescott, Edward C., "Should Control Theory Be Used for Economic Stabilization?" in Karl Brunner and Alan H. Meltzer, eds., *Optimal Policies, Control Theory, and Technology Exports*, Vol. 7, Carnegie-Rochester Conference on Public Policy Series, 1977, 13–38.

Rich, Robert W., "The Rationality of Price and Inflation Forecasts: New Evidence on the Livingston and the SRC Survey Series," Ch. 1 unpublished doctoral dissertation, Brown University, 1987.

Runkle, David E., "Liquidity Constraints and the Permanent Income Hypothesis: Evidence from Panel Data," manuscript, Federal Reserve Bank of Minneapolis, 1989.

Sargent, Thomas J., "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics," *Journal of Political Economy*, June 1976, *84*, 631–40.

Urich, Thomas and Wachtel, Paul, "The Structure of Expectations of the Weekly Money Supply Announcement," *Journal of Monetary Economics*, March 1984, *13*, 183–94.

Watson, Mark W., "Imperfect Information and Wage Inertia in the Business Cycle: A Comment," *Journal of Political Economy*, October 1983, *91*, 876–79.

West, Kenneth D., "Asymptotic Normality, When Regressors Have a Unit Root," *Econometrica*, November 1988, *56*, 1397–1417.

Zarnowitz, Victor, "An Appraisal of Short-Term Economic Forecasts," NBER Occasional Paper 104, New York: Columbia University Press. 1967.

______, "The New ASA-NBER Survey of Forecasts by Economic Statisticians," *American Statistician*, February 1969, *23*, 12–16.

______, "How Accurate Have the Forecasts Been?" in William F. Butler, Robert A. Kavesh, and Robert B. Platt, eds., *Methods and Techniques of Business Cycle Forecasting*, Englewood Cliffs, NJ: Prentice-Hall, 1974, pp. 565–96.

______, "Business Cycle Analysis and Expectational Survey Data," NBER Working Paper 1378, June 1984.

______, "Rational Expectations and Macroeconomic Forecasts," *Journal of Business and Economic Statistics*, October 1985, *3*, 293–311.

U.S. Department of Commerce, Bureau of Economic Analysis, *Business Conditions Digest*, various issues.

______, *Survey of Current Business*, various issues.

Margin Requirements, Volatility, and the Transitory Component of Stock Prices

By GIKAS A. HARDOUVELIS*

Official margin requirements in the U.S. stock market were established in October 1934 to limit the amount of credit available for the purpose of buying stocks. Since then, higher or rising margin requirements are associated with lower stock price volatility, lower excess volatility, and smaller deviations of stock prices from their fundamental values. The results hold throughout the post-1934 period and are not very sensitive to the exclusion of the turbulent depression years from the sample. Thus margin requirements seem to be an effective policy tool in curbing destabilizing speculation. (JEL 313, 520).

Federal regulation of securities margins is mandated by Congress in the Securities cchange Act of 1934. The stock market perience of the late 1920s led Congress to e conclusion that credit-financed specula- on in the stock market might create exces- 'e market volatility through a pyramiding- pyramiding process.[1] Congress reasoned at the imposition of margin requirements uld constrain the amount of borrowing d prevent excessive market volatility, and bsequently gave the Federal Reserve ju- sdiction over the level of margin require- ents. Since 1934, the Federal Reserve anged the initial margin requirements in stocks twenty-three times, and the current rate is 50 percent (see Table 1).[2]

Do initial margin requirements curb speculative excesses in the stock market and reduce stock price volatility? This question has recently gained new importance among regulators and students of financial market developments following the October 1987

*Department of Finance, Rutgers University, New nswick, NJ 08903; and Research Department, Fed- l Reserve Bank of New York, NY 10045. I would to thank the seminar participants at U. C. Berke- , Boston College, Columbia, Ohio State, Rutgers, Securities and Exchange Commission, the Central nk of Greece, and the Federal Reserve Bank of New rk, as well as Paul Bennett, Arturo Estrella, Ken- h Froot, Steve Peristiani, James Poterba, Anthony drigues, and George Sofianos for helpful discussions comments; and Valerie Laporte for editorial assis- ce. Special thanks go to James O'Brien, Robert ller, and two anonymous referees for very thorough nments. The views expressed in this article do not ect the views of the Federal Reserve Bank of New rk or the Federal Reserve System.

[1] The pyramiding-depyramiding process is reviewed Kenneth Garbade (1982) and can be summarized as ows: In the absence of broker-enforced margin re- rements, optimistic investors with relatively low de- es of risk aversion might borrow large amounts of funds to buy stock, causing a price rise that could not be justified by economic fundamentals. The price rise might then feed on itself if these speculators used their increased wealth to buy more stocks, thus driving prices even higher. This pyramiding effect could be followed by a market collapse if after an initial price drop, brokers and other creditors were to ask for more collateral on their loans to speculators. If some speculators could not provide the additional collateral, creditors would sell the stocks they kept as collateral, driving prices still lower. This outcome would generate further calls for collateral, more liquidations, and additional price declines.

[2] An initial margin requirement of, say, 60 percent implies that at least 60 percent of the value of the new stock purchase should come from the investor's own capital. If the stock price rises after the initial purchase, investors can withdraw the differential from their margin account or can use it to buy additional stock on 60 percent margin. If the price declines after the initial purchase, investors are not required to add funds to their margin accounts unless their equity position falls below the so-called maintenance margin, which has remained at 25 percent since 1934. Brokers typically add a spread over the official maintenance margin that varies across customers and across time. See George Sofianos (1988) for a detailed description of contemporary margin regulation in the cash and derivative markets.

TABLE 1—INITIAL MARGIN REQUIREMENTS

Effective	Rate (Percent)	Effective	Rate (Percent)
10/15/34	45	01/16/58	50
02/01/36	55	08/05/58	70
11/01/37	40	10/16/58	90
02/05/45	50	07/28/60	70
07/05/45	75	07/10/62	50[a]
01/21/46	100	11/06/63	70
02/01/47	75	06/08/68	80
03/30/49	50	05/06/70	65
01/17/51	75	12/06/71	55
02/20/53	50	11/24/72	65
01/04/55	60	01/03/74	50
04/23/55	70		

Sources: New York Stock Exchange *Fact Book*, 1987, p. 54; Board of Governors of the Federal Reserve System, *Annual Report*, various issues.
[a]Rate incorrectly reported in *Fact Book*.

abrupt collapse of stock prices and of the market mechanism.[3] Clearly, theory alone cannot provide a definite answer. Those who believe that speculation is stabilizing because it deepens the market and increases liquidity are likely to view margin requirements as harmful. Those who believe that an unchecked market is often subject to destabilizing speculation are likely to think that margin requirements could prevent speculative excesses. The question can only be resolved empirically.

Surprisingly, there is little, if any, previous empirical work that focuses on the role of margin requirements in curbing speculative excesses. Earlier researchers have concentrated primarily on the effect of a change in margin requirements on the level of stock prices and have found that an increase in margin requirements tends to stop and sometimes reverse a previous run-up in prices.[4] Stock price volatility is first mentioned by Thomas Moore (1966), who argues that margin requirements are an ineffective tool for controlling volatility. He observes that although margin requirements have changed a number of times since 1934, the volatility of stock prices has remained relatively stable. James O'Brien (1984) discusses the central issue of speculative excesses but foregoes an empirical investigation on the grounds that short-term speculative excesses have not been a characteristic of the post-1929 period. A detailed study by the Board of Governors of the Federal Reserve System (1984) takes a similar position, making no attempt to correlate a measure of stock price volatility or a measure of speculative excess with the size of margin requirements. Two timely articles in the literature—one by George Douglas (1969) and another by R. Officer (1973)—do find a negative association between the level of margin requirements and the volatility of stock prices. But these articles fall short of a complete analysis of the effects of margin requirements on stock volatility and do not examine excess volatility, the key variable of interest.

This paper conducts an empirical investigation of the role of initial margin requirements in curbing speculative excesses in the cash market. Section I discusses the theoretical connection between margin requirements and destabilizing speculation. Section II confirms the previous evidence on the negative correlation between stock price volatility and the level of margin requirements. The analysis exploits all the available data and controls for many excluded variables (particularly those associated with the Fed's regulatory response to the economic environment) that could have caused a bias in the results of Douglas and Officer. The robustness of the negative association between margin requirements and volatility is checked using two measures of volatility, one annual and the other monthly, plus a vector autoregressive (VAR) analysis. The VAR analysis shows that an increase in margin requirements is associated with a future drop in (1) stock return volatility, (2) stock returns, (3) the growth of trading volume at the New York Stock Exchange (NYSE), and (4) the deflated amount of

[3]See, for example, the "Interim Report of the Working Group on Financial Markets," submitted to the President of the United States, May 1988. See also Franklin Edwards (1988) and Arturo Estrella (1988).

[4]See Jacob Cohen (1966); Walter Eckardt and Donald Rogoff (1976); Corwin Grube, Maurice Joy, and Don Panton (1979); James Largay (1973); or James Largay and Richard West (1973). Dudley Luckett (1982) finds that margin requirements negatively affect the ratio of debt to equity in broker-dealer customer accounts.

borrowing for the purpose of buying stocks. Thus the evidence is consistent with the hypothesis that margin requirements constrain the activities of investors whose behavior raises market volatility.

Section III focuses on the topic of principal interest, the relationship between *excess* volatility and margin requirements. It employs a regression test of excess volatility that is robust to common errors of misspecification and allows for a time-varying discount rate. The test affirms the previous evidence on excess volatility and finds that excess volatility is weaker both during periods of high and during periods of rising margin requirements. The natural question that follows is which type of excessive market behavior do margin requirements seem to reduce? By construction, the test of Section III excludes the popular rational speculative bubbles hypothesis as an explanation of the excess volatility results. Hence, to answer the question, one has to pursue hypotheses of irrational behavior, and there is no well-accepted benchmark model for such behavior.

Section IV explores forms of irrational behavior that give rise to transitory components in stock prices (see Robert Shiller, 1984, or Lawrence Summers, 1986). One implication of the existence of transitory components or "fads" is the presence of negative correlation between price-dividend ratios and subsequent multiperiod stock returns. Another implication is the presence of negative serial correlation in multiperiod stock returns. I examine both implications and find that the transitory components in stock prices become less pronounced during periods of high and periods of rising margin requirements.

Section V offers some concluding remarks. An appendix contains the detailed exposition of the excess volatility test of Section III and a description of the data.

I. Margin Requirements and Destabilizing Speculation: Theoretical Linkages

The proposition that margin requirements help curb destabilizing speculation is based on two implicit claims: The first claim is that speculation by some groups of investors can be destabilizing. The second claim is that margin requirements can impose an effective constraint on the market activities of speculators. The first claim is plausible but is not shared by all economists. For example, Milton Friedman (1953) argues that speculation is destabilizing only if speculators on the average lose money by selling when assets are low in price and buying when assets are high. There are two counterarguments to Friedman's position. First, in a dynamic model with speculators that enter and exit the market, a few speculators can destabilize the market and subsequently perish. Second, even in a static context with perfect competition, it is possible to construct theoretical models of destabilizing speculation featuring speculators that do not necessarily lose money on the average (see Oliver Hart and David Kreps, 1986; Bradford DeLong, Andrei Shleifer, Lawrence Summers, and Robert Waldman, 1987; or John Campbell and Peter Kyle, 1988). These models show that speculation *can* destabilize prices, not that it necessarily destabilizes prices. Theory cannot predict the effect of increased speculation on the volatility of stock prices.

The claim that margin requirements can impose a binding constraint on the behavior of destabilizing speculators is also plausible. Finance theory predicts that the less risk-averse investors hold more stocks and less cash in their portfolios and are therefore more likely to be constrained by the imposition of margin requirements than are other investors. In practice, the margin requirement would not be an effective constraint if speculators could sell other assets they own or somehow find a way around the margin restriction at no extra cost, or if corporations could alter their financing behavior in order to attain the desired debt-equity ratio of their stockholders. Thus the importance of margin requirements as an effective constraint is an empirical question. In the empirical analysis of Sections III and IV we allow for nonlinearities arising from the possibility that margins may be more binding during the times when margin requirements change.

II. Margin Requirements and Volatility

This section contains an analysis of the relation between margin requirements and volatility. The volatility of stock returns, if measured properly, can capture unusual swings in stock prices caused by the pyramiding-depyramiding effect that prompted Congress to establish margin rules some 50 years ago. Since the pyramiding-depyramiding process is likely to last a few months, the volatility measure should be based on horizons that span a properly long time period. I employ an annual measure of volatility, but for robustness, I also use a monthly measure. Measures of volatility that are based on daily fluctuations of stock returns are not suited for analyzing the pyramiding-depyramiding process.

The section begins with a description of the two alternative measures of volatility. Then it investigates the economic variables that determine the Fed's propensity to change margin requirements. Finally, it presents multiple regression and vector autoregression analyses of the relation between margin requirements and volatility.

A. *Measures of Volatility*

Let r_t denote the real rate of return from the last trading day of month $t-1$ to the last trading day of month t. The variable r_t is the nominal return of the Standard & Poor's index including dividends minus the CPI inflation rate. To construct an annual measure of volatility, I first run the following regression:

$$(1)\quad r_t = \sum_{i=1}^{12} \alpha_i SEASON_{it} + \sum_{j=1}^{12} \beta_j r_{t-j} + \varepsilon_t,$$

where $SEASON_i$ is a monthly intercept. Then I define the annual measure of volatility, $\sigma_{y,T}$, as the standard deviation of the estimated residuals $\hat{\varepsilon}_t$ over a calendar year:

$$(2a)\quad \sigma_{y,T} \equiv \left[\sum_{j=1}^{12} \left(\hat{\varepsilon}_{T,j} - \hat{\varepsilon}_T\right)^2 / 11 \right]^{1/2},$$

where the index T denotes calendar years, the index j runs over the months of a calendar year, and $\hat{\varepsilon}_T$ is the average residual over the calendar year. The above definition of volatility allows the conditional mean of monthly stock returns to vary over time and across months. However, it assumes that volatility itself is constant over the calendar year. This assumption is not problematic because here the aim is to find a good empirical measure of long and large swings in stock prices and not to arrive at the best empirical proxy of volatility per se.[5]

Margin requirements do not conveniently change at the end of a calendar year, and hence σ_y does not match well with the margin requirement. For this reason and in order to check the robustness of the results, I also use G. William Schwert's (1988a) monthly measure of volatility:

$$(2b)\quad \sigma_{m,t} \equiv (\Pi/2)^{1/2} |\hat{\varepsilon}_t|,$$

where $|\hat{\varepsilon}_t|$ is the absolute value of the estimated residuals of regression equation (1). Under the hypothesis that the residuals in equation (1) are normally distributed, the adjustment factor $(\Pi/2)^{1/2}$ in (2b)—which is approximately equal to 1.2533—makes $\sigma_{m,t}$ an unbiased statistic of the true standard deviation. The variable σ_m is not a smooth measure, but for our purposes it suffices that it captures all the unusual spikes in monthly returns.[6]

Table 2 presents summary statistics of the two volatility measures for the period following the establishment of official margin

[5] An alternative annual measure of volatility would be the simple standard deviation of the monthly returns. This measure provides very similar results. Volatility can also be based on monthly excess nominal rates of return, that is, *ex-post* nominal returns minus the known one-month Treasury bill rate at the beginning of the one-month holding period. Again, this measure is not very different because the volatility of nominal stock prices overwhelms both the volatility of the CPI and the volatility of the Treasury bill rate.

[6] Note that the variable of interest in this paper is *actual* volatility, not perceived volatility. Schwert (1988b) constructs a measure of perceived monthly volatility from the fitted value of the regression of $\sigma_{m,t}$ on $\sigma_{m,t-j}$, r_{t-j}, and $SEASON_{jt}$ for $j = 1,\ldots,12$.

TABLE 2—VOLATILITY OF REAL STOCK RETURNS-SUMMARY STATISTICS

	σ_y	σ_m
Sample	1935–1987	1934:11–1987:12
Observations	53	638
Mean	0.044	0.045
Standard Deviation	0.019	0.040
Skewness	1.78	2.31
Kurtosis	4.51	9.69
Minimum	0.018	0.000
Maximum	0.120	0.332
ρ_1	0.44	0.11
ρ_2	0.21	0.14
ρ_3	0.04	0.18
ρ_{12}	0.01	0.13
ρ_{24}	–	−0.02
ρ_{36}	–	0.01
ρ_{48}	–	0.01
Rank Correlation	−0.34	−0.12
t-statistic	[−2.60]	[−3.07]

Notes: σ_y is the standard deviation from January to December of the estimated residuals $\hat{\varepsilon}_t$ from the regression: $r_t = \Sigma_{i=1}^{12}\alpha_i SEASON_{it} + \Sigma_{i=1}^{12}\beta_i r_{t-i} + \varepsilon_t$, where r_t is the monthly real rate of return of the S&P index including dividends, and *SEASON* is a monthly dummy variable. σ_m equals $(\pi/2)^{1/2}|\hat{\varepsilon}_t|$ and is a monthly measure of volatility. $\rho_1, \rho_2, \ldots, \rho_{48}$ denote the autocorrelation at lags $1,2,\ldots,48$. The rank correlation is between the margin requirement of month t and σ_m, and between the average margin during a calendar year and σ_y.

requirements. Both σ_y and σ_m are nonnormal and autocorrelated, and σ_m is more variable than σ_y. The table also provides a preview of the association between volatility and margin requirements. Given the nonnormality of both measures, rank correlations are computed instead of simple correlations. In both cases the rank correlation between volatility and margin requirements is negative. The *t*-statistics of the rank correlations are quite large but are only partly meaningful because they do not take into account the serial dependence of volatility.

Figure 1 plots σ_m together with the official margin requirement from 1927 through 1987. The figure shows that in the late twenties and early thirties, when the official margin requirement was zero, volatility was unusually high. However, during this early period of zero official margin requirements, brokers imposed substantial margins, and the zero official margin underestimated the true effective margin more severely than official margin of later periods did. Thus the inclusion of the late twenties and early thirties in the sample period biases the results in favor of finding a negative association between margin requirements and volatility. For this reason, the subsequent empirical analysis uses only the period when official margin requirements were in effect (November 1934 to the present). Naturally, the results are now biased against finding a negative association because a high volatility–low margin period carries zero weight in the regressions.

Figure 1 brings out a second important point. Because margin requirements have changed very few times, it would be extremely difficult to uncover a potential relation to volatility. Indeed, no such relation is obvious from the figure, especially during the postdepression years. Uncovering a potential relation between the two variables requires careful analysis and appropriate control for third factors that affect volatility.

B. *The Federal Reserve's Regulatory Behavior*

One factor complicating the empirical analysis is the Federal Reserve's reaction function. The Fed's behavior in regulating margins may have resulted in a spurious negative or positive correlation between margin requirements and volatility. This would be the case, for example, if the Fed had altered margin requirements in response to an economic variable that had itself been responsible for the change in volatility or in response to the change in volatility itself.

The Fed's official reasons for changing margin requirements are recorded in the various issues of the *Annual Report* of the Board of Governors of the Federal Reserve System. The Fed typically attributes its decision to increase margin requirements to a rapid increase in stock prices and suspected speculation and to a rapid expansion in stock market credit. Sometimes high trading volumes, inflationary pressures, and an expanding economy were also given as reasons

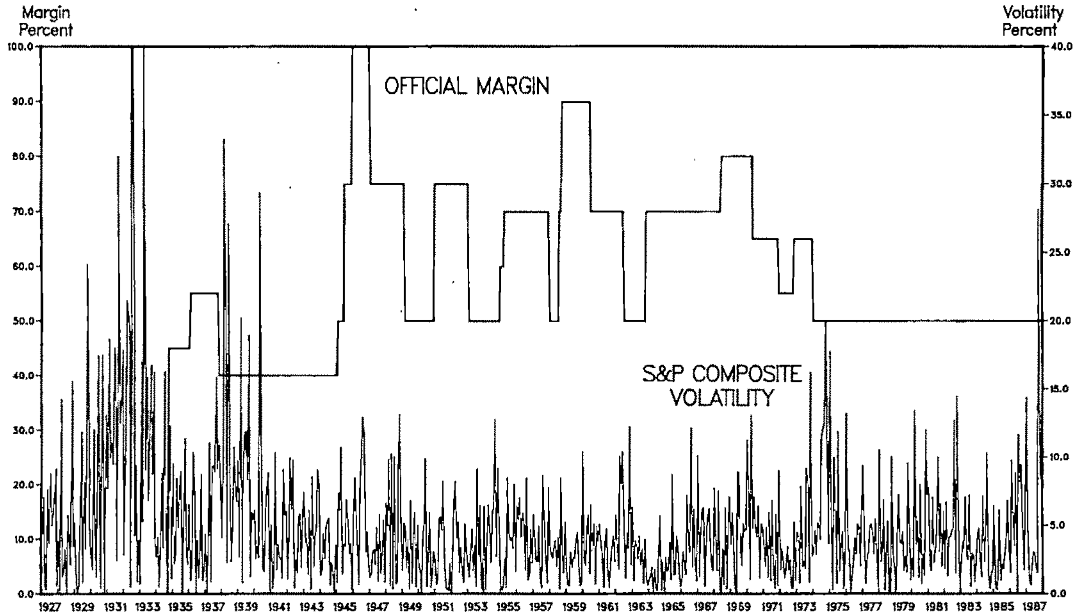


FIGURE 1. OFFICIAL MARGIN REQUIREMENT AND STOCK VOLATILITY

for an increase in margins. In deciding to decrease margin requirements, Fed officials usually cited a drop in stock market credit and the disappearance of those factors that had led to an earlier increase. Stock market volatility was never mentioned as a reason for changing margin requirements. The following excerpt from the 1951 *Annual Report* is characteristic of the explanations supplied by the Fed after an increase in margin requirements:

> Although the total amount of credit in use in the stock market had not assumed heavy proportions, there had been some increase during the preceding months together with increases in the volume of trading and in prices of securities. The expanding business and economic situation appeared to be encouraging stock market activity and speculation, and the Board of Governors believed that in the existing circumstances a further substantial price advance supported by a rapid expansion in stock market credit was a distinct possibility. The increase in margin requirements was effected as a preventive measure. [p. 81]

Table 3 presents estimates of a hypothetical Fed decision rule to change margin requirements using an ordered-response logit model with three states. Briefly, assume that the Fed's unobserved disposition to alter margin requirements, Z_t, has a logistic distribution. Assume also that a decrease in margin requirements occurs when $Z_t < \beta' X_t + c_1$, that no change in margin requirements occurs when $\beta' X_t + c_1 < Z_t < \beta' X_t + c_2$, and that an increase in margin requirements occurs when $\beta' X_t + c_2 < Z_t$. X_t represents a set of indicator variables and β', c_1, and c_2 are parameters to be estimated. The likelihood function to be maximized is

$$(3)\quad \prod_{t=1}^{T} F(\beta' X_t + c_1) \times [F(\beta' X_t + c_2) - F(\beta' X_t + c_1)] \times [1 - F(\beta' X_t + c_2)],$$

where F denotes the logistic cumulative distribution function.[7] The model is estimated

[7] See G. S. Maddala (1983, pp. 46–49) for more details. The maximization of the likelihood function was performed using *GAUSS*.

TABLE 3—THE FEDERAL RESERVE'S REACTION FUNCTION: ORDERED-RESPONSE LOGIT NOVEMBER 1934–DECEMBER 1987

Annual Horizon			Monthly Horizon		
Variable	Coefficient	*T*-Statistic	Variable	Coefficient	*T*-Statistic
c_1	-3.96	-2.91	c_1	-4.26	-8.40
c_2	5.60	3.34	c_2	4.53	8.81
$\bar{r}_{t-1}$	79.11	2.70	r_{t-1}	17.34	2.10
$\overline{\mathrm{MCR}}_{t-1}$	42.90	1.73	MCR_{t-1}	15.27	2.39
$\sigma_{y,t-1}$	-0.57	-0.02	$\sigma_{m,t-1}$	-1.63	-0.21
$\bar{\pi}_{t-1}$	119.35	0.72	π_{t-1}	52.73	1.00
$\bar{Y}_{t-1}$	-48.75	-1.55	Y_{t-1}	12.06	0.77
$\overline{VL}_{t-1}$	7.52	1.03	VL_{t-1}	0.25	0.20
Observations	638			638	
-2 ln(Likelihood)	191.20			200.67	

r_{t-1} ≡ real rate of stock return from end of month $t-2$ to end of month $t-1$ (nominal return including dividends minus CPI inflation rate).

$\bar{r}_{t-1}$ ≡ average of $r_{t-12}, r_{t-11}, \dots, r_{t-1}$.

MCR_{t-1} ≡ growth rate of the ratio of broker-dealer credit to customers over the capitalized value of the NYSE from month $t-2$ to month $t-1$.

$\overline{\mathrm{MCR}}_{t-1}$ ≡ average of $MCR_{t-12}, MCR_{t-11}, \dots, MCR_{t-1}$.

π_{t-1} ≡ CPI inflation rate from month $t-2$ to month $t-1$.

$\bar{\pi}_{t-1}$ ≡ average of $\pi_{t-12}, \pi_{t-11}, \dots, \pi_{t-1}$.

Y_{t-1} ≡ growth rate of the industrial production index from month $t-2$ to month $t-1$.

$\bar{Y}_{t-1}$ ≡ average of $Y_{t-12}, Y_{t-11}, \dots, Y_{t-1}$.

VL_{t-1} ≡ growth rate of the NYSE trading volume from month $t-2$ to month $t-1$.

$\overline{VL}_{t-1}$ ≡ average of $VL_{t-12}, VL_{t-11}, \dots, VL_{t-1}$.

$\sigma_{y,t-1}$ ≡ see equation (2a) of the text.

$\sigma_{m,t-1}$ ≡ see equation (2b) of the text.

The dependent variable takes the value of 1, 2, or 3 depending on whether the official margin decreased, stayed the same, or increased during month t.

over the sample period from November 1934 through December 1987, but the results are similar when we include the first change in official margin requirements of October 1934. The included indicator variables are the lagged real rate of return of stock prices, the lagged percentage change in stock market credit (defined as the dollar amount of dealer and broker credit to customers for the purposes of buying stocks deflated by the capitalized value of the NYSE), the lagged stock return volatility measures of Table 2, the lagged rate of inflation, the lagged growth in the industrial production index, and the lagged growth in trading volume at the NYSE. With the exception of volatility, all these variables were mentioned by the Fed as reasons for changing margin requirements. The model is estimated twice, once using the previous month's values of the indicator variables and σ_m as the measure of stock market volatility, and a second time using the previous year's average monthly values and σ_y as the measure of stock market volatility. In the latter estimation, σ_y is defined over a rolling twelve-month interval that ends one month before the observation of the dependent variable.

The results show that lagged stock returns and the lagged rate of changes in stock market credit were the only statistically significant variables influencing the Fed's decision to change margin requirements. An increase in the growth of stock prices (i.e., the level of stock returns) or in the growth of stock market credit raises the probability of an increase in margin requirements. As expected, volatility did not affect the decision to change margin requirements.[8]

If the Fed's response to a recent growth in stock prices is effective and the market subsequently declines, it will be difficult to uncover a negative association between margin requirements and stock market volatility. The reason is that volatility is low in bull markets and high in bear markets. Thus, after the Fed raises the margin requirement and the market declines, the bear market raises volatility and creates a spurious positive relation between the change in margin requirements and the change in volatility. Hence it is important to control for the effect of a rising or declining market in our subsequent investigation. The negative relationship between the level of stock prices and stock price volatility is analyzed extensively by Andrew Christie (1982). Christie also provides evidence supporting Fischer Black's (1976) hypothesis that the low debt-equity ratios during periods with high stock prices cause volatility to be low.

C. *Evidence from Regression Equations*

Let us turn now to the effect of margin requirements on stock price volatility, an issue analyzed previously by Douglas (1969) and Officer (1973). Douglas controls only for the variation in dividends in his regression equation, and Officer for the variation in the industrial production index. Neither paper provides a complete analysis because each paper's reference to margin requirements is only incidental. Here I run a more complete regression that includes additional factors affecting volatility and that uses all available observations on volatility. The additional variables include the following: first, the two variables to which the Fed was found to respond, namely, the lagged level of stock returns and the lagged growth of stock market credit; second, the average level of inflation and the average growth of the industrial production index. These variables capture the effect of the economic environment on volatility and avoid a possible spurious correlation between margin requirements and volatility. For example, in addition to changing margin requirements, the Fed may change monetary policy. The new monetary policy may affect the economy, and the state of the economy may in turn affect volatility. In such a case a negative correlation between margin requirements and volatility may arise not because the change in margins was effective but because the simultaneous change in monetary policy affected volatility. Third, the volatility of the growth rate of the industrial production index serving as a proxy for the volatility of dividends.[9]

The regression equations are estimated using Ordinary Least Squares, but I use the method of Lars Hansen (1982) with the weighting scheme of Whitney Newey and Kenneth West (1987) to adjust the coefficient standard errors for conditional heteroskedasticity and the presence of a moving average (MA) process in the error term. The autocorrelations of the regression residuals determine the choice of the lag length in the MA correction.

Table 4A uses nonoverlapping annual observations and relates $\sigma_{y,T}$ to the average official margin requirement of the calendar year, $\bar{M}_T$. Two sets of regressions are reported, one in which the margin requirement is the only explanatory variable and a second set with all the control variables added. The estimation is performed over a variety of subsamples in order to check the sensitivity of the results to particular subperiods. Figure 1 suggests that a negative asso-

[8] The results are qualitatively similar when I use the level of stock prices relative to trend instead of the level of stock returns. Also, binary logit models show that the decision to increase margins is symmetric to the decision to decrease margins (margin credit is slightly more significant in the case of a decrease in margins).

[9] In Hardouvelis (1988c), I include lagged volatility to control for the possibility—despite the evidence of Table 3—that the Fed responds to lagged volatility. The inclusion of lagged volatility creates an errors-in-variables problem, yet the results are very similar to the present results.

TABLE 4A—MARGIN REQUIREMENTS AND VOLATILITY—ANNUAL OBSERVATIONS

$$\sigma_{y,t} = \beta_0 + \beta_1 \overline{M}_t + \beta_R \bar{r}_t + \beta_{MCR} \overline{MCR}_t + \beta_\pi \bar{\pi}_t + \beta_y \overline{Y}_t + \beta_{vy} \sigma(Y_{y,t}) + v_t$$

Sample	Nobs	β_m	β_R	β_{MCR}	β_π	β_y	β_{vy}	$\overline{R}^2$	*SEE*	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	q
1935–87	53	−0.048*						0.11	0.018	0.30	0.04	−0.15	−0.18	−0.03	1
		(0.024)													
		−0.078*	−0.377*	−0.546*	−1.52	−0.001	0.447	0.35	0.015	0.28	−0.13	−0.27	−0.19	−0.03	1
		(0.020)	(0.127)	(0.151)	(0.92)	(0.264)	(0.262)								
1947–87	41	−0.037*						0.07	0.013	0.10	−0.18	−0.19	−0.01	0.10	0
		(0.017)													
		−0.053*	−0.203	−0.296	0.156	−0.718*	0.125	0.26	0.012	0.08	−0.11	−0.23	0.00	−0.07	0
		(0.023)	(0.167)	(0.196)	(0.861)	(0.342)	(0.289)								
1935–81	47	−0.047						0.11	0.018	0.32	0.08	−0.12	−0.18	−0.08	1
		(0.026)													
		−0.075*	−0.358*	−0.510*	−1.40	0.002	0.514*	0.35	0.015	0.28	−0.08	−0.28	−0.20	0.01	1
		(0.021)	(0.134)	(0.165)	(0.96)	(0.222)	(0.232)								
1935–55	21	−0.042						0.04	0.023	0.37	0.16	−0.13	−0.26	−0.19	1
		(0.034)													
		−0.043	−0.588*	−0.749*	−4.90*	0.528*	0.640	0.51	0.017	0.18	−0.30	−0.16	−0.11	0.01	0
		(0.023)	(0.124)	(0.159)	(1.24)	(0.208)	(0.361)								
1956–87	32	−0.047*						0.11	0.014	0.16	−0.21	−0.26	−0.08	0.12	0
		(0.018)													
		−0.058*	−0.225	−0.188	−0.054	−1.11*	0.195	0.30	0.012	0.18	−0.18	−0.12	−0.07	−0.03	0
		(0.027)	(0.175)	(0.251)	(1.07)	(.481)	(0.413)								

Notes: σ_Y is the volatility of monthly real rates of return of the S&P index over a calendar year (see Table 2). $\overline{M}$, $\bar{r}$, $\overline{MCR}$, $\bar{\pi}$, and $\overline{Y}$ denote the average margin requirement, the average real rate of return of the S&P index, the average monthly growth of the ratio of margin debt to the value of the NYSE, the average rate of inflation, and the average growth in the industrial production index over a calendar year. $\sigma(Y_y)$ is the volatility of the monthly growth in the industrial production index over a calendar year defined in a manner similar to σ_y. Nobs denotes the number of observations, $\overline{R}^2$ the coefficient of determination adjusted for degrees of freedom, *SEE* the regression standard error, and $\rho_1,\dots,\rho_5$ the residual autocorrelations at lags 1,...,5. Newey-West (1987) corrected standard errors are in parentheses. The correction is for conditional heteroskedasticity and for a moving average of order q. Asterisk, *, denotes statistical significance at the 5 percent level.

TABLE 4B—MARGIN REQUIREMENTS AND VOLATILITY—MONTHLY OBSERVATIONS

$$\sigma_{m,t} = \Sigma_{i=1}^{12} \alpha_i SEASON_{it} + \beta_m M_t + \beta_r r_{t-1} + \beta_{MCR} MCR_{t-1} + \beta_\pi \pi_t + \beta_y Y_t + \beta_{vy} \sigma(Y_{m,t}) + u_t$$

Sample	Nobs	β_m	β_r	β_{MCR}	β_π	β_y	β_{vy}	$\overline{R}^2$	*SEE*	ρ_1	ρ_2	ρ_3	ρ_{12}	ρ_{24}	$\chi^2(11)$ $\alpha_1 = \cdots = \alpha_{12}$	q
1934:	638	−0.037*						0.02	0.040	0.11	0.12	0.16	0.09	−0.07	21.8*	12
11−87:12		(0.018)													[0.026]	
		−0.049*	−0.233*	−0.126*	0.087	−0.073	0.291*	0.08	0.039	0.04	0.05	0.12	0.08	−0.09	23.6*	12
		(0.016)	(0.035)	(0.044)	(0.366)	(0.092)	(0.085)								[0.015]	
1947:	492	−0.027*						0.02	0.034	0.03	0.12	0.11	0.01	−0.11	17.2	3
1−87:12		(0.013)													[0.102]	
		−0.034*	−0.204*	−0.088	0.047	−0.325*	0.077	0.07	0.033	−0.03	0.06	0.08	0.01	−0.11	22.7*	0
		(0.012)	(0.053)	(0.051)	(0.388)	(0.129)	(0.151)								[0.020]	
1934:	566	−0.039*						0.02	0.039	0.12	0.13	0.19	0.11	−0.08	24.7*	12
11−81:12		(0.019)													[0.010]	
		−0.049*	−0.231*	−0.136*	0.091	−0.072	0.274*	0.08	0.038	0.05	0.04	0.14	0.10	−0.11	26.1*	12
		(0.017)	(0.039)	(0.047)	(0.381)	(0.094)	(0.082)								[0.006]	
1934	254	−0.033						0.02	0.046	0.14	0.14	0.18	0.10	−0.09	28.4*	12
11−55:12		(0.027)													[0.003]	
		−0.045*	−0.232*	−0.118*	−0.159	−0.029	0.242*	0.07	0.045	0.08	0.05	0.13	0.08	−0.12	28.1*	12
		(0.022)	(0.042)	(0.051)	(0.417)	(0.106)	(0.094)								[0.003]	
1956:	384	−0.031*						0.03	0.035	0.05	0.11	0.14	0.01	−0.10	16.1	5
1−87:12		(0.015)													[0.138]	
		−0.031*	−0.209*	−0.095	0.634	−0.261	0.066	0.08	0.034	−0.03	0.05	0.10	0.02	−0.11	17.5	3
		(0.015)	(0.064)	(0.070)	(0.640)	(0.152)	(0.203)								[0.093]	

Notes: σ_m is a monthly measure of volatility of the monthly real rate of return of the S&P index (see Table 2). $SEASON_i$ is a monthly dummy intercept, M, r, MCR, π, and Y are the official margin requirement, the real rate of return of the S&P index, the growth in the ratio of margin debt to the value of the NYSE, the rate of inflation, and the growth in the industrial production index. $\sigma(Y_m)$ is a monthly measure of volatility for Y defined in a similar manner as σ_m. $\rho_1,\dots,\rho_{24}$ denote the residual autocorrelations at lags 1,...,24. Newey-West (1987) corrected standard errors for conditional heteroskedasticity and a moving average of order q are in parentheses. Significance levels are in brackets below each $\chi^2(11)$ statistic. Asterisk, *, denotes statistical significance at the 5 percent level.

ciation may be due entirely to the depression years. Thus I reestimate each equation using the usual postwar sample from 1947 to the present. I also partition the sample in the middle, at the end of 1955, so that eleven changes in margin requirements occur in the first half and eleven in the second half. Finally, I show the results for the 1935–81 sample in order to exclude the period of existence of a futures market in the S&P index. Margin requirements in the futures market are much lower than the

TABLE 4C—CHANGE IN MARGIN REQUIREMENTS AND CHANGE IN VOLATILITY
SAMPLE: 22 CHANGES IN MARGIN REQUIREMENTS, 1936–1974

$$\Delta\sigma_{y,t} = \alpha_0 + \alpha_m \Delta M_t + \alpha_r \Delta \bar{r}_t + \alpha_{MCR} \Delta \overline{MCR}_t + \alpha_\pi \Delta \bar{\pi}_t + \alpha_y \Delta \bar{y}_t + \alpha_{vy} \Delta\sigma(Y_{y,t}) + \varepsilon_t$$

α_m	α_r	α_{MCR}	α_π	α_y	α_{vy}	$\bar{R}^2$	SEE	Degrees of Freedom
−0.009 (0.028)					0.387 (0.319)	−0.02	0.021	19
−0.158* (0.060)	−0.070* (0.033)	−0.984* (0.344)	−0.367 (0.214)	0.041 (0.032)	0.408 (0.392)	0.17	0.019	15

Notes: Δ denotes the change of a variable from period $(t-12, t-1)$ to period $(t+1, t+12)$, where t is the month when a margin change occurs. $\Delta\sigma_{y,t}$, Δr_t, $\Delta\overline{MCR}_t$, $\Delta\bar{\pi}_t$, $\Delta\bar{y}_t$, and $\Delta\sigma(Y_{y,t})$ are defined this way. ΔM_t is the change in margin requirements during month t. Asterisk, *, denotes statistical significance at the 5 percent level.

post-1974 cash market margin of 50 percent, creating an effective margin requirement lower than 50 percent.

Table 4A shows a statistically significant negative relationship between margin requirements and volatility. The results are stronger when the control variables are included in the regressions, as was expected. For example, consistent with our earlier discussion, volatility is negatively associated with stock returns; thus, excluding stock returns would bias the correlation between margin requirements and volatility in the positive direction. The negative association between margin requirements and volatility is observed in *every* subperiod, which is quite impressive. After all, the variability of the official margin is minimal and checking for subperiod relationships is asking too much of the data.

Table 4B repeats the analysis of Table 4A but uses monthly observations and relates the monthly measure of volatility, $\sigma_{m,t}$, to the margin requirement M_t. The regressions also include twelve separate monthly intercepts. The hypothesis that these intercepts are equal is rejected. The overall results are very similar to the results of Table 4A, an outcome that is reassuring. Margin requirements show a significant negative relation to volatility, and this negative relation is present in every subperiod.

The overall results are weaker than the findings of Douglas but stronger than those of Officer. Douglas relates the level of volatility to the level of margin requirements but uses four-year averages of these variables and looks at a cross section of stocks. Officer relates the change in volatility to the change in margin requirements. His sample begins before the establishment of official margins and ends in 1970. Since Officer finds a weaker relation than I do, it is interesting to uncover the source of our differences. Table 4C repeats Officer's exercise with my data and the post-1935 sample period. The table isolates the 22 cases since 1935 when a change in the official margin requirement took place, and regresses the change in the annual measure of volatility from the year spanning months $t-12$ to $t-1$ to the year spanning months $t+1$ to $t+12$ on the change in the margin requirements of month t. The first row of Table 4C contains the results of the regression used by Officer: the only explanatory variables are the change in the margin requirement and the change in the volatility of the monthly growth of the industrial production index. The results are weaker than Officer's.[10] The second row adds the extra proper controls and shows that, indeed, margin requirements have a very strong and statistically significant effect on volatility. Clearly, Officer's weak results were due to misspecification. He excluded factors that affect both volatility and margin requirements.

[10] Officer does not describe his timing very well, and it is possible that he may have examined the difference between the volatility that spans months $t-5$ to $t+6$ and the volatility that spans months $t-6$ to $t+5$. Also, the R^2 coefficients in his regressions are very high, which is puzzling.

TABLE 4D—MARGIN REQUIREMENTS AND VOLATILITY-VECTOR AUTOREGRESSIONS
SAMPLE: 1935:10–1987:12
INDEPENDENT VARIABLES: 12 LAGS OF $\sigma_{m,t}$, M_t, r_t, AND MCR_t PLUS 12 MONTHLY INTERCEPTS

Dependent Variable	Independent Variables									
	M		σ_m		r		MCR			
	SUM	$\chi^2(12)$	*SUM*	$\chi^2(12)$	*SUM*	$\chi^2(12)$	*SUM*	$\chi^2(12)$	$\bar{R}^2$	*SEE*
σ_m	−0.025*	20.4	0.55*	30.8*	−0.29*	50.3*	−0.31*	17.8	0.17	0.037
	(0.012)	[0.059]	(0.13)	[0.002]	(0.13)	[0.000]	(0.12)	[0.123]		
M	0.96*	10046*	−0.13	7.6	0.27*	16.8	0.14	13.4	0.94	0.034
	(0.012)	[0.000]	(0.07)	[0.819]	(0.11)	[0.159]	(0.09)	[0.337]		
r	−0.020	15.9	0.08	13.2	0.32	28.2*	−0.02	30.4*	0.06	0.046
	(0.015)	[0.195]	(0.15)	[0.351]	(0.17)	[0.005]	(0.15)	[0.002]		
MCR	−0.000	60.3*	−0.14	20.9	0.78*	128.2*	0.28*	32.0*	0.32	0.039
	(0.013)	[0.000]	(0.11)	[0.051]	(0.15)	[0.000]	(0.14)	[0.001]		

Notes: σ_m, M, r, and MCR are defined in Table 3. *SUM* denotes the sum of the coefficients of the twelve lags of each independent variable. White (1980) heteroskedasticity-consistent standard errors are in parentheses. The $\chi^2(12)$ statistic tests that the coefficients of the 12 lags are jointly zero. Significance levels are in brackets. Asterisk, *, denotes statistical significance at the 5 percent level.

D. *Reduced Form Evidence: Vector Autoregressions*

The negative relation between margin requirements and volatility does not necessarily imply causation from margin requirements to volatility. However, the opposite causation from volatility to margin requirements is not supported by the evidence and is not intuitive. As I mentioned earlier, there is not a single instance when the Fed cited volatility even as a remote cause for its decision to change margin requirements. Also, the ordered-logit results of Table 3 show no such relation. Furthermore, suppose for the sake of the argument that the Fed observed an increase in volatility and became concerned about excessive speculative activity. Under such a scenario, the Fed would ordinarily increase margin requirements; only a decision to decrease the requirements would justify the negative association between the two variables. Despite these arguments, Schwert (1988b) criticizes an earlier version of this paper on the grounds that the negative association between the two variables reflects the effects of volatility on margin requirements. Schwert regresses volatility on twelve of its own lags and on twelve lags and twelve leads of the change in the official margin. Finding that lags of margin changes have a small negative effect while leads of margin changes have a large negative effect on volatility, he concludes that margin requirements do not affect volatility but volatility affects margin requirements. However, Schwert's conclusion is hasty. Further examination shows that the link from lagged volatility to current margin requirements reflects third factors.

Table 4D presents multivariate autoregressions with four variables and twelve lags. The variables are the level of margin requirements, the monthly measure of stock volatility, the monthly real rate of return of stocks, and the growth in the ratio of margin credit to the value of the NYSE. Thus Table 4D adds the two control variables that are significant in Tables 3 and 4B to the usual bivariate relation between margin requirements and volatility. Observe that the Schwert results are now reversed: There is no effect from volatility to margin requirements, but there is an effect from margin requirements to volatility. The table also suggests an explanation for Schwert's counterintuitive results: Consistent with our earlier discussion, stock returns affect volatility negatively and quite strongly, and margin requirements positively. Thus the negative effect of volatility on margin requirements that Schwert finds may simply reflect the positive effect of stock returns on margin

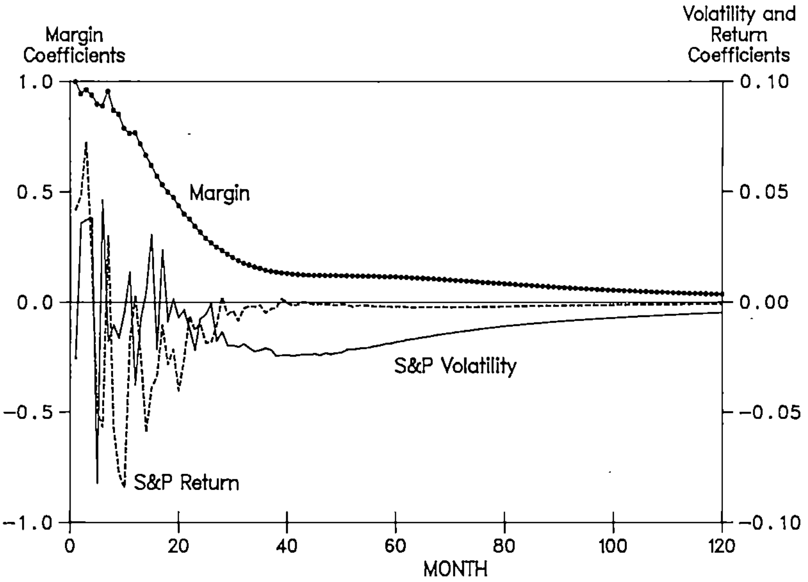


FIGURE 2. IMPULSE RESPONSE TO A POSITIVE SHOCK IN OFFICIAL MARGIN REQUIREMENTS

requirements. Of course, other influences are also present. The table shows multidirectional feedback between the variables—a finding that makes individual results hard to interpret. Also, the reduced form coefficients of a VAR model cannot be given a precise structural interpretation in general. In the case of margin requirements in particular, the VAR model treats margin requirements as a continuous variable that depends on its own past lags, which does not have a meaningful interpretation.

One way to visualize the dynamic effects of margin requirements on volatility and on other variables of interest is to trace through time the effects of a shock to margin requirements using a previously estimated VAR system. Such an exercise can provide a useful final summary of the direction of the various responses, although in the absence of standard errors, it cannot assess the strength of these responses. To perform the exercise, I added to the previous list two new variables: the volatility of the industrial production index and the growth in trading volume. The first variable was significant in the regressions of Tables 4A and 4B when the full sample was used. The second variable is an interesting output variable to examine. Computational limitations did not permit the inclusion of more variables.

Figure 2 presents the impulse response of stock volatility and real stock returns to a temporary shock in the official margin requirement. The margin requirements shock itself persists for about two years and dies out completely after ten years. The return to zero is a reflection of a mean-reverting tendency of the official margin during the sample period. Real stock returns respond negatively for about two years. Volatility also responds negatively and for a longer time. The size of the initial volatility response is comparable to the size of the regression coefficients in Table 4B. Note that most of the previous literature on margin requirements concentrated essentially on one of the two pieces of information in the figure—stock returns—and ignored volatility.

Figure 3 provides some more interesting evidence. It shows the impulse response of margin credit and trading volume. Both variables fall after an increase in margin requirements. Thus taken together, Figures 2 and 3 provide evidence consistent with the

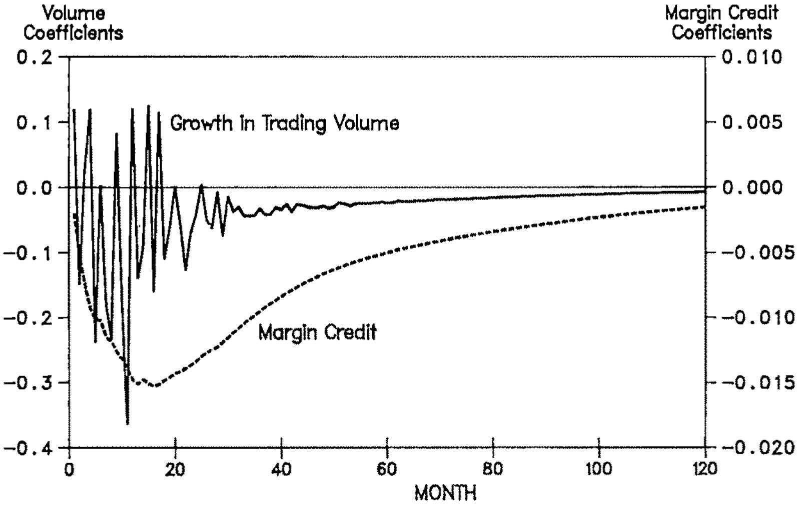


FIGURE 3. IMPULSE RESPONSE TO A POSITIVE SHOCK IN OFFICIAL MARGIN REQUIREMENTS

Note: Margin credit is the ratio of the broker-dealer credit to customers over the capitalized value of the NYSE.

hypothesis that margin requirements constrain the activity of investors whose behavior raises market volatility.[11]

III. Margin Requirements and Excess Volatility

The previous section showed that margin requirements are negatively correlated with stock price volatility. However, volatility per se is not a direct measure of speculative excess. A better measure is excess volatility, defined as volatility that cannot be accounted for by the variation of current and future dividends and discount rates. Although the regressions of Tables 4A, 4B, and 4C controlled for the variation of fundamental determinants of stock prices, they did not impose the present value relation on the data and thus cannot be interpreted as referring to excess volatility. This section investigates the relation between margin requirements and excess volatility, employing a regression-based test used previously by Louis Scott (1985) and Kenneth Froot (1987).

A. *Theoretical Framework*

The test of excess volatility is based on a very common test for the rationality of expectations, the so-called unbiasedness test. Consider a variable y, and let f denote an earlier forecast of y and u the corresponding forecast error: $y_{t+1} \equiv f_t + u_{t+1}$. The variable f is a *rational* (optimal) forecast of y based on publicly available information if the forecast error u is uncorrelated with the forecast f. The unbiasedness test can be performed by running the regression

$$(4) \qquad u_{t+1} = \beta_0 + \beta_1 f_t + \varepsilon_{t+1},$$

and testing the hypothesis that $\beta_0 = 0$ and $\beta_1 = 0$. Froot (1987) observed that the nature of the rejection of $\beta_1 = 0$ has an interesting interpretation: $\beta_1 < 0$ implies *excess* volatility, while $\beta_1 > 0$ implies *insufficient* volatility (see Appendix A for details). For

[11] I have repeated most of the empirical work of this section using an index of small stocks—the two smallest deciles of the NYSE in terms of capitalized value. The results for small stocks are *stronger* than the results for the S&P index and were reported in Hardouvelis (1988c). Some of those results can also be found in Hardouvelis (1988b).

my purposes, it is sufficient to show that the rejection of the rationality hypothesis is more severe in periods with low or decreasing margin requirements. Nevertheless, I adopt Froot's interpretation to claim that excess volatility worsens in those periods.

To apply the test to stock prices, I follow Kenneth West (1988) and Froot (1987) and fix investors' holding period horizon to, say, one period. Note that the period length can be varied arbitrarily for empirical purposes. For this holding period, the present value model implies that the stock price is the discounted present value of the expected nominal dividend, E_tD_{t+1}, plus the expected proceeds from selling the stock at the end of the period, E_tP_{t+1}:

$$(5)\quad P_t=(E_tP_{t+1}+E_tD_{t+1})/(1+E_tR_{t+1}),$$

where E_tR_{t+1} is the nominal discount rate or required rate of return. Put differently, equation (5) claims that the required rate of return equals the expected rate of return, ruling out opportunities for expected profits. In the test's framework, P_t serves as the forecasting variable f. The variable y would then be the present value of the *actual* proceeds of holding a stock for one period and selling it at the end of the period:

$$(6)\quad y_{t+1}\equiv(P_{t+1}+D_{t+1})/(1+E_tR_{t+1}).$$

Let us now express the required rate of return as the geometric sum of the observable nominal risk-free rate, i_t, and a risk premium, $E_t\theta_{t+1}$:

$$(7)\quad 1+E_tR_{t+1}\equiv(1+i_t)(1+E_t\theta_{t+1}).$$

The actual nominal rate of return, R_{t+1}, is as follows:

$$(8)\quad 1+R_{t+1}\equiv(P_{t+1}+D_{t+1})/P_t.$$

Identities (6), (7), and (8) lead to

$$(9)\quad y_{t+1}\equiv P_t(1+R_{t+1})/(1+i_t)(1+E_t\theta_{t+1})$$
$$\cong P_t(1+R_{t+1}-i_t-E_t\theta_{t+1}).$$

From equations (5) and (9), the forecast error u_{t+1} is

$$(10)\quad u_{t+1}\equiv y_{t+1}-P_t$$
$$\cong P_t(R_{t+1}-i_t-E_t\theta_{t+1}).$$

Using equation (10), one could proceed to regress u_{t+1} on P_t, as in equation (4). One problem with the test is the nonstationarity of the level of stock prices. Since P_t appears both on the right-hand side and the left-hand side of the regression of u_{t+1} on P_t and is nonstationary, the regression results become meaningless. This problem can be resolved, however, if we can find a variable x_t such that P_t/x_t is stationary. We can then regress u_{t+1}/x_t on P_t/x_t.

Suppose we find a nonzero correlation between u_{t+1} and P_t—that is, either excess or insufficient volatility. What are the economic reasons for such a phenomenon? Robert Flood, Robert Hodrick, and Paul Kaplan (1986) argue that one popular explanation in the literature, rational speculative bubbles, cannot account for such an empirical finding. A bubble component in a stock price, b_t, is rational if it is expected to grow at the required rate of return E_tR_{t+1} (see Hardouvelis, 1988a). Thus a rational bubble is of the form

$$b_{t+1}=(1+E_tR_{t+1})b_t+v_{t+1},$$
$$\text{with } E_tv_{t+1}=0,$$

where v_{t+1} is a white noise shock to the bubble. The presence of a rational bubble adds to the forecast error $u_{t+1}\equiv y_{t+1}-P_t$ the term

$$b_{t+1}/(1+E_tR_{t+1})-b_t$$
$$=v_{t+1}/(1+E_tR_{t+1}),$$

which is white noise and is uncorrelated with P_t.

B. *Empirical Evidence on Excess Volatility*

Let us begin by testing for deviations from rationality by regressing the forecast error u_{t+1} of equation (10) on the stock price P_t.

Following Scott (1985) and Froot (1987), I deflate the stock price by a measure of dividends to transform it into a stationary series. In our framework, the choice of dividends as the deflator series provides statistical power because deviations from fundamentals affect the price-dividend ratio. I use a smoothed dividend measure, a three-year rolling average. The price-smoothed dividend ratio has even more power to detect deviation from fundamentals. Economic theory claims that in the absence of bubbles, the price-dividend ratio (as well as the price-smoothed dividend ratio) *is* stationary. John Campbell and Robert Shiller (1987) argue that it is also stationary in practice. In order to test the null hypothesis of nonstationarity, I regress the first difference in P_t/d_t on its lagged level and on twelve lagged first differences over the sample period from 1930 to the present. The t-statistic of the lagged level is -3.2, which is significant at the 5 percent level (see Wayne A. Fuller, 1976, p. 373).[12]

The basic regression equation is as follows:

$$(11a)\quad (R_t - i_{t-k})(P_{t-k}/d_{t-k}) = \beta_0 + \beta_1(P_{t-k}/d_{t-k}) + \varepsilon_t,$$

where k is the number of months in the holding period horizon, i_{t-k} is a k-month yield to maturity of a Treasury security observed at the end of month $t-k$, P_{t-k} is the stock price at the end of month $t-k$, and d_{t-k} is the average dividend over the period from $t-k-35$ through $t-k$. Note that the risk premium, which is unobservable, is missing from the above regression equation. Thus the error term contains noise plus the term $E_{t-k}\theta_t(P_{t-k}/d_{t-k})$. This may bias the coefficient estimates, but we show later that the bias is against finding excess volatility. Note also that in (11a), the price-dividend ratio is not forecasting future excess returns but the product of excess returns with the price dividend ratio. In Section IV, I will analyze the price-dividend ratio as a predictor of future excess stock returns; the latter regression has a slightly different interpretation.

Table 5A presents the regression results of equation (11a) for holding period horizons that range from one month to five years. To check the sensitivity of the results to the inclusion of the turbulent depression years, equation (11a) is reestimated excluding the 1930s from the sample. The table presents both sets of estimates. The estimation uses nonoverlapping observations for the horizons of one month, three months, and one year, and overlapping annual observations for the remaining horizons.[13] All reported standard errors are corrected for conditional heteroskedasticity. In the horizons of two to five years, the procedure of Newey and West (1987) is used to correct the standard errors for the moving average generated by the overlapping of the data.

The Chi-squared statistic for $\beta_0 = 0$ and $\beta_1 = 0$ tests the joint null hypothesis of forecast rationality and a zero risk premium. As expected, this joint hypothesis is rejected. The slope coefficient β_1 is negative, indicating the presence of excess volatility (assuming the risk premium is constant). Surprisingly, the results for the postdepression sample show even stronger excess volatility. In the full sample, the one-month horizon estimate of β_1 is insignificant and is similar to the evidence of Froot.[14] The results for

[12] Schwert (1987) cannot reject the null hypothesis of nonstationarity in the postwar data. Similarly, I was unable to reject the null hypothesis for the price-smoothed dividend ratio in Schwert's 1947–87 sample or in the 1940–87 sample. The inability of the test to reject the null in these *smaller* samples is probably an indication of the test's low power (see John Cochrane, 1988, for related criticisms) and is not a problematic issue. The price-smoothed dividend ratio *is* more stationary than price itself. Put differently, a plot of the price-smoothed dividend ratio across time does not reveal any obvious explosive behavior, and for the purposes of this paper this result is all that matters.

[13] The results are similar when monthly overlapping observations are used. These results are available on request.

[14] Froot uses the CRSP data on the return of the NYSE index. He also reports results for a horizon of one year, but in the absence of appropriate data on the one-year risk-free rate, he rolls over one-month T-bill rates in order to construct excess one-year returns.

TABLE 5A—ARE STOCK PRICES EXCESSIVELY VOLATILE?

$$(R_t - i_{t-k})\left(\frac{P_{t-k}}{d_{t-k}}\right) = \beta_0 + \beta_1\left(\frac{P_{t-k}}{d_{t-k}}\right) + \varepsilon_t$$

Horizon	Sample	Nobs	β_0	β_1	$\bar{R}^2$	*SEE*	*DW*	H_0: $\beta_0 = \beta_1 = 0$
			Full Sample					
1 month	1929–87	707	0.41*	−0.012	0.00	1.25	1.81	9.67*
			(0.19)	(0.08)				[0.008]
3 months	1929–87	233	0.99*	−0.026	0.00	2.02	1.65	11.02*
			(0.43)	(0.018)				[0.004]
1 year	1930–87	58	5.30*	−0.160*	0.04	4.82	1.83	10.70*
			(1.78)	(0.070)				[0.005]
2 years	1931–87	57	11.80*	−0.371*	0.12	6.91	0.86	18.05*
			(3.14)	(0.130)				[0.000]
3 years	1932–87	56	16.53*	−0.497*	0.16	8.16	0.45	20.34*
			(4.49)	(0.191)				[0.000]
4 years	1933–87	55	21.89*	−0.623*	0.19	9.34	0.44	22.69*
			(5.45)	(0.223)				[0.000]
5 years	1934–87	54	19.34*	−0.367	0.03	12.40	0.75	15.49*
			(7.88)	(0.307)				[0.000]
			Postdepression Sample					
1 month	1940–87	575	0.57*	−0.017*	0.01	1.11	1.85	22.31*
			(0.18)	(0.008)				[0.000]
3 months	1940–87	189	1.46*	−0.040*	0.02	1.77	1.67	21.83*
			(0.44)	(0.018)				[0.000]
1 year	1941–87	47	6.63*	−0.189*	0.08	3.97	2.03	17.59*
			(1.98)	(0.076)				[0.000]
2 years	1942–87	46	13.41*	−0.376*	0.19	5.07	1.02	34.32*
			(3.22)	(0.128)				[0.000]
3 years	1943–87	45	20.83*	−0.576*	0.31	5.84	0.77	64.37*
			(4.17)	(0.179)				[0.000]
4 years	1944–87	44	28.35*	−0.766*	0.35	7.20	0.65	69.55*
			(5.28)	(0.230)				[0.000]
5 years	1945–87	43	30.78*	−0.710*	0.16	10.77	0.88	38.68*
			(7.90)	(0.334)				[0.000]

$$\Theta_{t-1}\left(\frac{P_{t-1}}{d_{t-1}}\right) = \beta_0 + \beta_1\left(\frac{P_{t-1}}{d_{t-1}}\right) + \varepsilon_t$$

	Sample	Nobs	β_0	β_1	$\bar{R}^2$	*SEE*	*DW*
Risk-Proxy 1	1929–87	706	0.77*	0.016*	0.02	1.05	2.03
			(0.16)	(0.006)			
	1929–87	706	Monthly	0.043*	−0.01	1.06	2.05
			Seasonals	(0.002)			
Risk-Proxy 2	1929–87	706	0.00	0.012*	0.23	0.03	1.88
			(0.001)	(0.001)			

Notes: R_t is the nominal return on stocks including dividends from the end of month $t-k$ to the end of month t. i_{t-k} is the yield to maturity of a k-month security at the end of month $t-k$. Returns are not annualized. P_{t-k} is the stock price at the end of month $t-k$ and d_{t-k} is the average dividend over the 36-month period before $t-k$. Θ_{t-1} is the risk premium at the end of month $t-1$. The 1-month, 3-month, and 1-year horizons use nonoverlapping observations. The 2- through 5-year horizons use overlapping annual observations. Corrected standard errors are inside the parentheses. The correction is for conditional heteroskedasticity and for the moving average generated by the overlapping data. $\bar{R}^2$ is the coefficient of determination adjusted for degrees of freedom, *SEE* is the regression standard error, and DW the Durbin-Watson statistic. * and # statistical significance at the 5 percent and 10 percent level in a two-tailed test. The last column shows chi-squared statistics with their significance levels in brackets. Risk-Proxy 1 is the absolute value of the residuals of the one-month horizon regression (multiplied by 1.2533); its two regressions are estimated using Cochrane-Orcutt. Risk-Proxy 2 is the spread between the annualized yields of Moody's Baa and Aaa corporate bonds at $t-1$, times P_{t-1}/d_{t-1}. This regression is estimated in first-differenced form.

the remaining horizons of three months and one to five years are significant and represent entirely new evidence. Scott's evidence refers to an even longer horizon and is similar, although he did not specifically interpret his rejection of the present value model as originating from excess volatility.

Table 5A also presents the results of regressing proxies of the excluded $E_{t-k}\theta_t(P_{t-k}/d_{t-k})$ term on the independent variable P_{t-k}/d_{t-k}. The first risk proxy is based on the volatility of the residuals of regression equation (11a) when the holding period is one month. The volatility measure is proportional to the absolute value of the residuals, but the results are similar when a rolling standard deviation is used instead.[15] For robustness, I also use a second risk proxy based on the product of the price-dividend ratio and the spread between the end-of-month Moody's Baa and Aaa long-term corporate bond yields. All regressions uniformly show that the excluded risk term is positively correlated with the independent variable, and thus its exclusion biases the estimated coefficients against finding excess volatility.

The use of market-determined multiperiod discount rates makes the excess volatility test of equation (11a) immune to the misspecification problems that arise when a single period discount rate is cumulated to construct a multiperiod discount rate. Flood, Hodrick, and Kaplan (1986) argue that if misspecification errors plague the single period discount rate, these errors cumulate in the constructed multiperiod discount rates and result in progressively stronger rejections of the present value model as the holding period horizon lengthens. They use this insight to claim that West's (1988) findings on excess volatility are due to such misspecification errors. They point out that rejections of the present value model grow stronger with the length of the holding period horizon in West's test, and take this as evidence that the cause of the rejection is misspecification rather than excess volatility. In contrast, the rejections of the present value model in Table 5A are uniform across holding period horizons. In Table 5A returns are not annualized and the estimated slope coefficients β_1 increase proportionately with the holding period horizon. Annualized returns would result in approximately equal slope coefficients.[16] In short, the result on excess volatility appears robust and cannot easily be attributed to an unknown misspecification error on the discount rate. Hence, a novel aspect of our test is the greater confidence it provides in tracing the rejection of the present value model to excess volatility.

C. *Do Margin Requirements Matter?*

To see the effects of margin requirements on excess volatility, I run the following regression:

$$(11\text{b})\quad (R_t - i_{t-k})(P_{t-k}/d_{t-k}) = \beta_0 + \beta_H H_t(P_{t-k}/d_{t-k}) + \beta_L L_t(P_{t-k}/d_{t-k}) + \varepsilon_t,$$

where H_t (L_t) is a dummy variable that takes the value of unity if the average margin requirement from the end of month $t-k$ to the end of month t is larger than (lower than or equal to) 50 percent. If margin requirements curb speculative excesses, excess volatility should be more prominent during periods of low margin requirements. One should then observe that the coefficient β_L is more negative than the coefficient β_H.

[15] Robert Malkiel (1975) and Robert Pindyck (1984), among others, claim that the volatility of stock prices and the equity premium are positively correlated. Robert Merton (1980) uses a similar working assumption. However, Kenneth French, William Schwert, and Robert Stambaugh (1987) find a very weak link between expected volatility and realized excess stock returns. As Poterba and Summers (1986) point out, the weak link may be due to simultaneity: There is opposite negative feedback from stock returns to volatility, an issue that we discussed extensively in Section II, Part B.

[16] The slopes line up closer to each other when all overlapping monthly observations are used. See Hardouvelis (1988c).

TABLE 5B—MARGIN REQUIREMENTS AND THE EXCESS VOLATILITY OF STOCK PRICES

$$\text{Panel A: } (R_t - i_{t-k})\left(\frac{P_{t-k}}{d_{t-k}}\right) = \beta_0 + \beta_H H_t\left(\frac{P_{t-k}}{d_{t-k}}\right) + \beta_L L_t\left(\frac{P_{t-k}}{d_{t-k}}\right) + u_t$$

$$\text{Panel B: } (R_t - i_t - k)\left(\frac{P_{t-k}}{d_{t-k}}\right) = \beta_0 + \beta_1\left(\frac{P_{t-k}}{d_{t-k}}\right) + \beta_p POS_t\left(\frac{P_{t-k}}{d_{t-k}}\right) + \beta_N NEG_t\left(\frac{P_{t-k}}{d_{t-k}}\right) + v_t$$

	Panel A					Panel B				
Horizon	β_H	β_L	$\overline{R}^2$	*SEE*	H_0: $\beta_H = \beta_L$	β_1	β_p	β_N	$\overline{R}^2$	*SEE*
					Full Sample					
1 month	−0.013	−0.016	0.00	1.25	0.24	−0.013#	0.022*	0.006	0.01	1.25
	(0.009)	(0.014)			[0.626]	(0.008)	(0.008)	(0.012)		
3 months	−0.026	−0.027	0.00	2.03	0.00	−0.028	0.040*	−0.032	0.02	2.01
	(0.019)	(0.028)			[0.959]	(0.018)	(0.016)	(0.025)		
1 year	−0.214*	−0.341*	0.11	4.65	4.60*	−0.179*	0.161*	−0.003	0.12	4.63
	(0.075)	(0.110)			[0.032]	(0.071)	(0.054)	(0.067)		
2 years	−0.460*	−0.745*	0.30	6.20	7.77*	−0.371*	0.240*	0.008	0.21	6.55
	(0.121)	(0.180)			[0.005]	(0.149)	(0.107)	(0.094)		
3 years	−0.599*	−0.986*	0.40	6.92	8.69*	−0.505*	0.239#	−0.002	0.22	7.86
	(0.167)	(0.231)			[0.003]	(0.243)	(0.133)	(0.132)		
4 years	−0.740*	−1.20*	0.43	7.83	10.73*	−0.576*	0.087	−0.068	0.18	9.40
	(0.183)	(0.212)			[0.001]	(0.284)	(0.116)	(0.143)		
5 years	−0.512#	−0.921*	0.13	11.73	4.67*	−0.661*	−0.074	−0.138	0.20	11.19
	(0.285)	(0.336)			[0.031]	(0.338)	(0.141)	(0.144)		
					Postdepression Sample					
1 month	−0.018#	−0.020	0.01	1.11	0.10	−0.018*	0.023*	0.009	0.01	1.10
	(0.009)	(0.014)			[0.750]	(0.008)	(0.009)	(0.011)		
3 months	−0.038*	−0.033	0.01	1.78	0.13	−0.043*	0.049*	−0.030	0.05	1.74
	(0.019)	(0.025)			[0.714]	(0.018)	(0.014)	(0.026)		
1 year	−0.231*	−0.297*	0.09	3.95	1.46	−0.224*	0.145*	0.023	0.16	3.78
	(0.082)	(0.114)			[0.227]	(0.077)	(0.048)	(0.058)		
2 years	−0.444*	−0.583*	0.24	4.91	4.20*	−0.346*	0.121	−0.046	0.25	4.90
	(0.118)	(0.140)			[0.040]	(0.118)	(0.084)	(0.063)		
3 years	−0.667*	−0.890*	0.40	5.42	9.83*	−0.506*	0.089	−0.108	0.37	5.60
	(0.155)	(0.156)			[0.002]	(0.180)	(0.091)	(0.086)		
4 years	−0.872*	−1.15*	0.43	6.71	10.04*	−0.661*	0.039	−0.129	0.36	7.13
	(0.202)	(0.207)			[0.002]	(0.264)	(0.098)	(0.121)		
5 years	−0.828*	−1.13*	0.20	10.49	3.86*	−0.868*	0.032	−0.129	0.38	8.76
	(0.304)	(0.342)			[0.049]	(0.342)	(0.131)	(0.142)		

Notes: See the notes of Table 5A. H_t (L_t) is a dummy variable that takes the value of unity if the average margin requirement from $t-k$ through t is greater than (less than or equal to) 50 percent. POS_t (NEG_t) is dummy variable that takes the value of unity if from $t-k$ to t the margin requirement increased (decreased). The column H_0: $\beta_H = \beta_L$ contains a chi-squared statistic with its significance level in brackets.

Panel A of Table 5B contains the regression results. Consistent with the hypothesis that margin requirements curb speculative excesses, β_L is more negative than β_H. Furthermore, the hypothesis that β_L equals β_H is rejected for all the holding period horizons from one to five years. These results do not depend on the low margin requirements of the depression years. They are present in the postdepression sample as well. Note that the dummy-variable methodology provides qualitative but not quantitative evidence on the effects of margin requirements on excess volatility. Given that margin requirements changed only 23 times, it would be very hard to extract robust quantitative information from the data, especially for horizons longer than one or three months.

An alternative way of partitioning the sample to see the effects of margin requirements is as follows:

$$(11c)\quad (R_t - i_{t-k})(P_{t-k}/d_{t-k}) = \beta_0 + \beta_1(P_{t-k}/d_{t-k}) + \beta_P POS_t(P_{t-k}/d_{t-k}) + \beta_N NEG_t(P_{t-k}/d_{t-k}) + \varepsilon_t,$$

where POS_t (NEG_t) is a dummy variable that takes the value of unity if the margin requirement increased (decreased) from the end of month $t-k$ to the end of month t. This formulation provides evidence of possible nonlinearities in the effect of margin requirements on excess volatility. For example, if investors find partial ways around the margin requirements constraint in the longer run, then during times when margin requirements increase, the constraint would be more binding than usual and excess volatility would be lower. One would also expect to find stronger evidence over the shorter horizons because such nonlinearities are not likely to last very long. Intuition also suggests that stronger nonlinear effects should be observed when margin constraints become more restrictive; that is, we ought to observe a strong β_P and a weaker β_N.

Panel B of Table 5B contains the regression results. Regression coefficient β_1 remains negative and significant. Regression coefficient β_P reflects the extra effect during periods when an increase in margins occurs. It is positive and significant for horizons up to two years, implying that excess volatility is lower during these times. The strength at short horizons is consistent with expectations. Regression coefficient β_N reflects the extra effect during periods when a decrease in margins occurs. As expected, it is negative but insignificant. The results are similar in the postdepression sample.

IV. Margin Requirements and the Transitory Component of Stock Prices

The previous section showed that excess volatility of stock prices becomes less pronounced during periods with high or increasing margin requirements. Since this excess volatility cannot be accounted for by rational bubbles, it is interesting to isolate the type of irrational behavior that margin requirements appear to mitigate. This section examines one type of irrational behavior, investment "fads."

The fads hypothesis claims that investors often act out of herd instinct and slowly drive prices away from fundamentals. Alternatively, noise traders with highly autocorrelated misperceptions may also drive prices slowly away from fundamentals. Because these deviations are small over short horizons and take a long time to die out, rational investors cannot arbitrage them away. Summers (1986) argues that deviations of this nature may be hard to detect even with hundreds of years of data. In his stylized example the stock price contains a transitory component with an autocorrelation coefficient close to unity. Such a transitory component is practically indistinguishable from a random walk, yet it generates long swings in stock prices away from fundamentals. Fama and French (1988a) observe that although irrational long swings in stock prices are hard to detect over the short run, they generate negative autocorrelation in multiperiod returns. In another paper, Fama and French (1988b) also contend that irrational long swings in stock prices generate a negative correlation between future multiperiod stock returns and current price-dividend ratios. The intuition for both correlations is straightforward: If the current stock price is irrationally high, it implies a high price-dividend ratio and a high current return. Later on, as the price slowly reverts to its fundamental value, multiperiod returns become negative; that is, they move in the opposite direction from current price-dividend ratios and current returns. Fama and French (1988a, b) present evidence of a negative correlation between these variables.[17]

[17] For tests of serial correlation in stock returns that are based on second moment statistics, see Andrew Lo and Craig McKinlay (1988), O'Brien (1984), or Poterba and Summers (1988). Related papers are by Bruce Lehman (1988) and Campbell and Shiller (1988).

Certainly, as previous authors emphasize, negative correlation between future multiperiod returns and either current price-dividend ratios or current multiperiod returns do not necessarily reject the market efficiency hypothesis. A negative correlation can also be generated by a model with time-varying risk premia as follows: An unusually high current stock price—one that generates high current price-dividend ratios and high current returns—may rationally reflect a low risk premium; and the low subsequent stock returns would simply reflect a low realized reward for the small amount of risk that market participants had correctly expected to assume. Thus, if high or rising margin requirements reduce the above negative correlations, one can conclude that margin requirements reduce either irrational price swings or the rationally perceived riskiness in stocks. Under either interpretation, however, margin requirements are effective: They reduce the size of transitory components and hence excess volatility, and/or they reduce the market's riskiness.

Before turning to the effect of margin requirements, let us replicate the previous evidence on the presence of a negative correlation between future returns and both current price-dividend ratios and current returns. I estimate the following regression equations:

$$(12a)\quad R_t - i_{t-k} = \beta_0 + \beta_1(P_{t-k}/d_{t-k}) + \varepsilon_t,$$

$$(13a)\quad r_t = \alpha_0 + \alpha_1 r_{t-k} + u_t,$$

where R_t is the realized nominal rate of return including dividends from the end of month $t-k$ to the end of month t, r_t is the realized real rate of return from $t-k$ to t (R_t minus the CPI inflation rate), k denotes the length of the forecasting horizon in months, and P_{t-k}/d_{t-k} is the price-smoothed dividend ratio at the end of month $t-k$. Note the similarity between equation (12a) and equation (11a) of Section III. In (11a) the dependent variable is the product of $R_t - i_{t-k}$ with P_{t-k}/d_{t-k}, not $R_t - i_{t-k}$ alone. Equation (11a) is derived from a basic definition of excess or insufficient volatility; a negative β_1 implies excess volatility. Equation (12a), however, is based on intuition, and a negative β_1 implies the presence of transitory components in stock prices that contribute to excess volatility. Turning to equation (13a), note that it is expressed in terms of real rates of return in order to accord with the formulation of earlier authors, but the results are similar when excess nominal rates are used instead (see Hardouvelis, 1988b).

Panels A and B of Table 6A present the results for equations (12a) and (13a), respectively. Panel A replicates the Fama and French (1988b) results: β_1 is negative and significant both in the full sample and in the postdepression sample. Panel B, however, finds very little evidence of a negative autocorrelation in multiperiod stock returns. Whatever negative autocorrelation exists is due to the depression years. The results are similar to those of O'Brien (1987), who also used the S&P index. O'Brien shows that the coefficients become significant only when the sample is extended back to the 1870s. Poterba and Summers (1988) justify the lack of a strong negative autocorrelation by appealing to the test's low power.[18]

To see if margin requirements have an effect on the correlations of Table 6A, I follow the same procedure as in equations (11b) and (11c) of Section III. I begin by dividing the sample period into regimes of high and low margin requirements. I run the following regression equations:

$$(12b)\quad R_t - i_{t-k} = \beta_0 + \beta_H H_t(P_{t-k}/d_{t-k}) + \beta_L L_t(P_{t-k}/d_{t-k}) + \varepsilon_t,$$

$$(13b)\quad r_t = \alpha_v + \alpha_H H_t r_{t-k} + \alpha_L L_t r_{t-k} + u_t,$$

[18] In Hardouvelis (1988c), I show that small stock returns exhibit stronger negative autocorrelations over longer horizons, at three, four, and five years. The relative strength of the coefficients of small stocks in the multiperiod return regressions is consistent with the evidence of Fama and French (1988a), who rank the NYSE stocks by size into portfolio deciles.

TABLE 6A—IS THERE A TRANSITORY COMPONENT IN STOCK PRICES?

$$\text{Panel A: } R_t - i_{t-k} = \beta_0 + \beta_1 \frac{P_{t-k}}{d_{t-k}} + \varepsilon_t$$

$$\text{Panel B: } \quad r_t = \alpha_0 + \alpha_1 r_{t-k} + u_t$$

	Panel A						Panel B					
Horizon	Sample	Nobs	$\beta_1 \times 100$	$\bar{R}^2$	*SEE*	*DW*	Sample	Nobs	α_1	$\bar{R}^2$	*SEE*	*DW*
						Full Sample						
1 month	1929–87	707	−0.086# (0.049)	0.01	0.060	1.78	1926–87	743	0.103 (0.070)	0.01	0.059	2.61
3 months	1929–87	233	−0.176* (0.0888)	0.02	0.089	1.83	1926–87	246	−0.087 (0.102)	0.01	0.124	1.99
1 year	1930–87	58	−1.17* (0.38)	0.13	0.215	1.99	1928–87	60	0.007 (0.140)	−0.02	0.224	1.95
2 years	1931–87	57	−2.47* (0.50)	0.27	0.295	1.20	1930–87	58	−0.177# (0.097)	0.02	0.312	1.22
3 years	1932–87	56	−3.35* (0.77)	0.35	0.339	0.59	1932–87	56	−0.268 (0.170)	0.07	0.364	0.56
4 years	1933–87	55	−4.79* (0.97)	0.42	0.426	0.93	1934–87	54	−0.152 (0.202)	0.01	0.474	0.78
5 years	1934–87	54	−3.62* (1.36)	0.16	0.599	1.08	1936–87	52	−0.016 (0.179)	−0.02	0.540	0.66
						Postdepression Sample						
1 month	1940–87	575	−0.089* (0.026)	0.02	0.042	1.91	1940–87	575	0.061 (0.051)	0.00	0.043	1.99
3 months	1940–87	189	−0.237* (0.075)	0.04	0.071	1.76	1940–87	191	0.192* (0.077)	0.03	0.079	1.89
1 year	1941–87	47	−1.16* (0.39)	0.17	0.168	2.06	1941–87	47	0.064 (0.168)	−0.02	0.169	1.92
2 years	1942–87	46	−2.36* (0.53)	0.35	0.217	1.07	1942–87	46	−0.109 (0.114)	−0.01	0.260	0.93
3 years	1943–87	45	−3.76* (0.73)	0.51	0.254	0.88	1943–87	45	0.070 (0.206)	−0.02	0.333	0.74
4 years	1944–87	44	−5.19* (0.95)	0.55	0.325	1.01	1944–87	44	0.248 (0.185)	0.04	0.403	0.65
5 years	1945–87	43	−5.42* (1.32)	0.42	0.437	0.99	1945–87	43	0.199 (0.145)	0.02	0.501	0.54

Notes: R_t, i_{t-k}, P_{t-k}, and d_{t-k} are defined in Table 5A. r_t is the real rate of return of the S&P-500 index from period $t-k$ through period t, defined as the nominal sock return R_t minus the CPI rate of inflation. See the notes of Table 5A for remaining definitions. Nonoverlapping observations are used for horizons up to 1 year. Overlapping annual observations are used for horizons 2 through 5 years. Inside the parentheses are standard errors corrected for conditional heteroskedasticity and for the moving average created by the data overlapping.

where H_t (L_t) is a dummy variable that takes the value of unity when the average initial margin requirement from $t-k$ through t is larger than (lower than or equal to) 50 percent. Panel A of Table 6B shows the results of equation (12b). β_H is less negative than β_L, which suggests that transitory components are less pronounced in periods of high margin requirements. The hypothesis that $\beta_H = \beta_L$ is rejected. This results holds in the full sample and in the postdepression sample as well; thus it is not due to the depression years alone. Panel B of Table 6B shows the results for equation (13b). In the full sample, there is strong evidence that α_H is not as negative as α_L. The hypothesis that $\alpha_H = \alpha_L$ is rejected at multiperiod horizons. This result, however, is entirely due to the depression years. Contrary to the results for β_H and β_L, there is no evidence of a difference between α_H and α_L in the postdepression sample.

TABLE 6B—MARGIN REQUIREMENTS AND THE TRANSITORY COMPONENT OF STOCK PRICE

Panel A: $R_t - i_{t-k} = \beta_0 + \beta_H H_t \left(\frac{P_{t-k}}{d_{t-k}} \right) + \beta_L L_t \left(\frac{P_{t-k}}{d_{t-k}} \right) + u_t$

Panel B: $r_t = \alpha_0 + \alpha_H H_t r_{t-k} + \alpha_L L_t r_{t-k} + V_t$

	Panel A					Panel B				
Horizon	$\beta_H \times 100$	$\beta_L \times 100$	$\bar{R}^2$	*SEE*	H_0: $\beta_H = \beta_L$	α_H	α_L	$\bar{R}^2$	*SEE*	H_0: $\alpha_H = \alpha_L$
					Full Sample					
1 month	−0.092#	−0.109	0.01	0.060	0.59	0.108	0.101	0.01	0.059	0.00
	(0.054)	(0.070)			[0.443]	(0.067)	(0.084)			[0.948]
3 months	−0.186*	−0.207#	0.01	0.089	0.21	0.199*	−0.221#	0.04	0.122	9.07*
	(0.093)	(0.166)			[0.647]	(0.084)	(0.132)			[0.003]
1 year	−1.43*	−2.03*	0.20	0.206	7.21*	−0.024	0.028	−0.03	0.226	0.04
	(0.37)	(0.49)			[0.007]	(0.188)	(0.192)			[0.838]
2 years	−2.88*	−4.19*	0.44	0.259	13.55*	0.005	−0.303*	0.04	0.309	3.20#
	(0.41)	(0.58)			[0.000]	(0.145)	(0.104)			[0.073]
3 years	−3.79*	−5.50*	0.56	0.280	17.06*	0.023	−0.486*	0.15	0.347	5.51*
	(0.61)	(0.74)			[0.000]	(0.217)	(0.130)			[0.019]
4 years	−5.33*	−7.44*	0.59	0.355	20.32*	0.163	−0.445*	0.13	0.445	9.75*
	(0.71)	(0.81)			[0.000]	(0.208)	(0.168)			[0.002]
5 years	−4.37*	−6.46*	0.26	0.562	7.49*	0.171	−0.388*	0.06	0.519	21.14*
	(1.29)	(1.64)			[0.006]	(0.143)	(0.118)			[0.000]
					Postdepression Sample					
1 month	−0.093*	−0.100*	0.02	0.042	0.18	0.098	0.039	0.00	0.043	0.40
	(0.029)	(0.041)			[0.671]	(0.065)	(0.070)			[0.529]
3 months	−0.240*	−0.244*	0.04	0.071	0.01	0.186*	0.241	0.03	0.079	0.06
	(0.82)	(0.108)			[0.917]	(0.082)	(0.211)			[0.807]
1 year	−1.40*	−1.77*	0.19	0.165	2.73#	0.013	0.170	−0.04	0.170	0.28
	(0.42)	(0.56)			[0.099]	(0.194)	(0.263)			[0.598]
2 years	−2.75*	−3.55*	0.43	0.203	7.12*	−0.057	−1.02*	0.01	0.258	7.33*
	(0.49)	(0.63)			[0.008]	(0.124)	(0.328)			[0.007]
3 years	−4.26*	−5.48*	0.62	0.223	15.83*	0.004	0.567	−0.01	0.333	1.41
	(0.59)	(0.68)			[0.000]	(0.218)	(0.435)			[0.236]
4 years	−5.75*	−7.20*	0.64	0.292	14.81*	0.203	0.768	0.03	0.404	0.79
	(0.77)	(0.86)			[0.000]	(0.208)	(0.539)			[0.374]
5 years	−6.09*	−7.84*	0.50	0.409	7.95*	0.213	0.038	−0.01	0.507	0.30
	(1.13)	(1.47)			[0.005]	(0.146)	(0.343)			[0.586]

Note: See the definitions in Tables 5A, 5B, and 6A.

Let us now turn to the possibility that margin requirements may be particularly influential at times when they change, especially when they become more constraining. I estimate the following regression equations:

$$(12c)\quad R_t - i_{t-k} = \beta_0 + \beta_1 (P_{t-k}/d_{t-k}) + \beta_P POS_t (P_{t-k}/d_{t-k}) + \beta_N NEG_t (P_{t-k}/d_{t-k}) + \varepsilon_t,$$

$$(13c)\quad r_t = \alpha_0 + \alpha_1 r_{t-k} + \alpha_P POS_t r_{t-k} + \alpha_N NEG_t r_{t-k} + u_t,$$

where POS_t (NEG_t) is a dummy variable that takes the value of unity when the level of initial margin requirements increases (decreases) from the end of month $t-k$ to the end of month t. Panel A of Table 6C contains the regression results of equation (12c). Observe that β_1 continues to be negative and significant. The interesting finding is that β_P is positive and significant, suggesting that transitory components are less pronounced during times that margin requirements increase. This is true both in the full sample and in the postdepression sample. β_N is not different from zero, implying the absence of nonlinearities when margin re-

TABLE 6C—MARGIN REQUIREMENTS AND THE TRANSITORY COMPONENT OF STOCK PRICES

Panel A: $R_t - i_{t-k} = \beta_0 + \beta_1\left(\frac{P_{t-k}}{d_{t-k}}\right) + \beta_p POS_t\left(\frac{P_{t-k}}{d_{t-k}}\right) + \beta_N NEG_t\left(\frac{P_{t-k}}{d_{t-k}}\right) + \varepsilon_t$

Panel B: $r_t = \alpha_0 + \alpha_1 r_{t-k} + \alpha_P POS_t r_{t-k} + \alpha_N NEG_t r_{t-k} + u_t$

	Panel A					Panel B				
Horizon	$\beta_1 \times 100$	$\beta_p \times 100$	$\beta_N \times 100$	$\bar{R}^2$	*SEE*	α_1	α_p	α_N	$\bar{R}^2$	*SEE*
				Full Sample						
1 month	−0.088#	0.077*	0.013	0.01	0.060	0.099	0.374*	0.158	0.01	0.059
	(0.049)	(0.031)	(0.043)			(0.072)	(0.150)	(0.359)		
3 months	−0.182#	0.160*	−0.118	0.03	0.088	−0.111	0.516*	0.496	0.01	0.124
	(0.108)	(0.061)	(0.087)			(0.105)	(0.175)	(0.385)		
1 year	−1.26*	0.66*	0.01	0.18	0.209	0.02	0.080	−0.538	−0.02	0.225
	(0.38)	(0.23)	(0.23)			(0.179)	(0.269)	(0.520)		
2 years	−2.52*	1.06*	0.12	0.35	0.280	−0.253*	0.748*	−0.137	0.15	0.290
	(0.58)	(0.42)	(0.34)			(0.125)	(0.246)	(0.165)		
3 years	−3.52*	1.21*	0.20	0.42	0.321	−0.466*	0.282	0.332	0.07	0.363
	(0.91)	(0.53)	(0.48)			(0.185)	(0.261)	(0.252)		
4 years	−4.94*	0.62	0.15	0.41	0.430	−0.059	−0.261	−0.057	−0.02	0.480
	(1.24)	(0.45)	(0.55)			(0.223)	(0.296)	(0.243)		
5 years	−5.74*	0.01	−0.06	0.43	0.494	0.063	0.052	−0.210	−0.04	0.545
	(1.43)	(0.69)	(0.56)			(0.177)	(0.214)	(0.175)		
				Postdepression Sample						
1 month	−0.092*	0.085*	0.029	0.02	0.042	0.059	0.464*	−0.128	0.00	0.043
	(0.026)	(0.031)	(0.040)			(0.052)	(0.186)	(0.439)		
3 months	−0.246*	0.191*	−0.104	0.07	0.070	0.187*	0.325	−0.136	0.03	0.079
	(0.073)	(0.062)	(0.087)			(0.084)	(0.228)	(0.200)		
1 year	−1.30*	0.60*	0.08	0.24	0.161	−0.075	0.617*	0.069	0.01	0.166
	(0.38)	(0.18)	(0.20)			(0.200)	(0.267)	(0.378)		
2 years	−2.32*	0.63#	−0.07	0.40	0.209	−0.096	0.659*	−0.240	0.10	0.246
	(0.52)	(0.34)	(0.24)			(0.167)	(0.271)	(0.243)		
3 years	−3.71*	0.60	−0.15	0.54	0.247	0.117	−0.066	−0.054	−0.07	0.341
	(0.83)	(0.41)	(0.37)			(0.277)	(0.269)	(0.356)		
4 years	−5.32*	0.56	0.03	0.55	0.326	0.370*	−0.269	−0.089	0.01	0.409
	(1.27)	(0.43)	(0.55)			(0.148)	(0.182)	(0.323)		
5 years	−6.46*	0.58	0.06	0.60	0.364	0.200	0.044	−0.049	−0.03	0.513
	(1.49)	(0.57)	(0.59)			(0.152)	(0.198)	(0.174)		

Note: See the definitions in Tables 5A, 5B, and 6A.

quirements are relaxed. Panel B of Table 6C shows the results of estimating equation (13c). Here the evidence is stronger than in Panel B of Table 6B, where I separated regimes of high and low margin requirements. An increase in margin requirements has a nonlinear effect, which is present across the whole sample. The coefficient α_P is positive and significant at short horizons both in the full sample and in the postdepression sample.

Overall, the evidence from the price-smoothed dividend ratios shows that the presence of transitory components is mitigated both during times that margin requirement are high and during times that they increase. The strength of the evidence from the autocorrelations of multiperiod stock returns is much weaker, however, and depends primarily on the influence of the depression years. In judging the strength of the latter results, one should recall the claim made by Poterba and Summers that tests of the null hypothesis of no serial correlation in returns have low power against the alternative of fads. In view of the low power, the lack of overwhelming statistical significance is not very surprising. Yet observe that despite their low power, autocorrelations do detect nonlinear effects that are also present in the postdepression sample.

V. Conclusion

Despite the limitations imposed on the analysis by the infrequent changes in the official margin requirements, the paper uncovered an interesting empirical regularity: Higher initial margin requirements in the cash market are associated with a reduction in both actual and excess stock price volatility and with a reduction in the size of long-run stock price swings away from fundamental values. The results are observed across the full sample and are not very sensitive to the exclusion of the turbulent depression years.

The case for interpreting the link between the level of official margin requirements and market speculative excesses in a causal way is strengthened by the presence of nonlinear effects at times when margin requirements increase. During those times, margin requirements apparently impose a more severe binding constraint on the behavior of destabilizing speculators and hence decrease stock price deviations away from fundamental values and reduce excess volatility. I conclude that, consistent with the aim of the Security and Exchange Act of 1934, initial margin requirements in the cash market seem to mitigate destabilizing speculation.

The results of the paper do not necessarily argue for active Fed intervention on a month-to-month basis in order to control volatility at high frequencies. The core finding is that margin requirements dampen *long* swings in stock prices that increase volatility at low frequencies and that originate from phenomena such as fads, the pyramiding and depyramiding of stock prices. A minimum policy prescription from the results is that policymakers should refrain from eliminating official margin requirements in cash equity markets. Even when margin requirements are kept constant for a long period of time, they could be effective in diminishing destabilizing speculative activity. The usual intuition that sophisticated investors are able to find ways around the margin restriction at a very small extra cost in the long run does not apply to every investor. In an evolving market with investors entering and exiting every period, margin requirements can be quite effective in constraining the activities of new entrants in the market.[19]

Since the stock market crash of October 1987, the role of derivative markets in index-based contracts has become a major topic in the public policy debate. Futures and options markets in stock indices are praised for providing liquidity and hedging capabilities to large institutional investors, but the same markets are also suspected of creating excessive volatility that spills over to the cash market.[20] To date, margin requirements in derivative markets are much lower than margin requirements in the cash market. The reason is that the primary aim of margins in derivative equity markets has been to reduce the probability of contractual defaults and the risk of a derivative market breakdown, assuming that stock price volatility is a given exogenous factor. The cash market experience of the last fifty years suggests, however, that long-term volatility is not exogenous and could be affected adversely by the low margin requirements in these markets.

APPENDIX A: TESTING FOR EXCESS VOLATILITY

Let σ_i^2 denote the variance of a variable i, ρ the correlation between u and f, and q the correlation between y and f. Following Froot (1987), let us take a mean-preserving spread around the forecasting variable f, keeping σ_y^2 and q constant. Then, if the mean squared prediction error σ_u^2 increases (decreases), f is

[19] Before making any stronger policy recommendations, it is desirable to examine the effects of *contemporary* margin policy in other countries. In the U.S. margin requirements have not changed since 1974, but in other countries, such as Japan, margin requirements are actively administered today. In a detailed study of the Japanese margin experience, Hardouvelis and Peristiani (1989–90) report that during the 1980s an increase in margin requirements had an economically and statistically significant negative impact on the volatility and momentum of Japanese stock prices.

[20] See Lawrence Harris (1988).

defined to be an excessively (insufficiently) volatile forecast of y. This definition leads to the definition in the text as follows: First, note that $q\sigma_f\sigma_y \equiv Cov(y,f) = \sigma_f^2 + \rho\sigma_f\sigma_u$, and thus $q\sigma_y = \sigma_f + \rho\sigma_u$. Next, $d\sigma_u^2/d\sigma_f = 2(\sigma_f - q\sigma_y)$. Substituting $\sigma_f + \rho\sigma_u$ for $q\sigma_y$, $d\sigma_u^2/d\sigma_f$ simplifies to $-2\rho\sigma_u$. Thus a negative (positive) correlation between u and f implies excessive (insufficient) forecast volatility. If we have data on y and f, we can easily determine whether f is a rational, an excessively volatile, or an insufficiently volatile forecast of y by running the regression $u = \beta_0 + \beta_1 f_t + \varepsilon$ and checking to see whether β_1 is zero, negative, or positive respectively.

How does the voluminous literature on volatility tests relate to the regression test above? As Froot (1987) discusses in detail, in this literature f is the stock price and y is the perfect foresight price. Volatility tests can be interpreted as rejecting the rationality hypothesis that the correlation coefficient ρ between the forecast f and the forecast error u is zero (Scott, 1985). If u and f are uncorrelated, then $\sigma_y^2 = \sigma_f^2 + \sigma_u^2$, and hence $\sigma_y^2 > \sigma_f^2$. The volatility literature uses direct estimates of σ_y^2 and σ_f^2 and finds that the previous inequality is violated. That is, empirically,

$$\sigma_y^2 < \sigma_f^2 \leftrightarrow \sigma_u^2 + 2\rho\sigma_u\sigma_f < 0$$

$$\leftrightarrow \text{plim}\,\beta_1 \equiv \rho\sigma_u/\sigma_f < -1/2(\sigma_u/\sigma_f)^2.$$

Thus in terms of the unbiasedness test (4) of the text, volatility tests have indirectly found that the slope coefficient β_1 is significantly smaller than $-1/2(\sigma_u/\sigma_t)^2$. Clearly, the unbiasedness test (4) is more general in discovering excess volatility, for it only requires that β_1 be smaller than zero. Next, note that a second volatility test of the rationality hypothesis is to test the inequality $\sigma_y^2 > \sigma_u^2$. In terms of our unbiasedness test (4), violation of this inequality is equivalent to

$$\sigma_y^2 < \sigma_u^2 \leftrightarrow \sigma_f^2 + 2\rho\sigma_u\sigma_f < 0$$

$$\leftrightarrow \text{plim}\,\beta_1 \equiv \rho\sigma_u/\sigma_f < -1/2.$$

Again, β_1 significantly smaller than $-1/2$ is a more stringent condition than β_1 significantly smaller than zero.

APPENDIX B: DATA AND SOURCES

The S&P Composite data are drawn from the 1988 yearbook of the Ibbotson Associates and represent end-of-month values from 1926 through 1987. The data come in two forms: with and without reinvested dividends, a distinction that allows the construction of a monthly dividend series. Currently, the index includes 500 of the largest stocks, but before March 1957 it consisted of 90 of the largest stocks.

The source for data on the one-month T-bill yield to maturity is also Ibbotson Associates. Data on the three-month and one- to five-year Treasury yields to maturity come from Stephen Cecchetti (1988) for the period January 1929 through December 1946, and from Houston McCulloch in Shiller and McCulloch (1987) for the period January 1947 through February 1987. The McCulloch data refer to zero coupon yields. Finally, the end-of-month Moody's Aaa and Baa corporate bond yields of Table 5A are drawn directly from Moody's Investor Services.

Data on the Consumer Price Index are taken from Ibbotson Associates, and on the industrial production index from the following sources: (i) for the period 1926–1946, *Industrial Production*, 1986 edition, Board of Governors of the Federal Reserve System; (ii) for the period 1947–October 1987, Citibase data banks; (iii) for November and December 1987, *International Financial Statistics*, April 1988 edition.

Sources for data on broker and dealer margin credit are (i) the series entitled "Customer Net Debit Balances" in Board of Governors of the Federal Reserve System, *Banking and Monetary Statistics*, 1943, Table no. 143; and *Banking and Monetary Statistics: 1941–1970*, 1976, Table no. 12.23; and (ii) the series entitled "Credit Extended to Margin Customers," which appears in various issues of the *Federal Reserve Bulletin* under the "Stock Market Credit" table. The first series runs from November 1931 through June 1978, and the second series runs from March 1967 through December 1987. The two series are not identical. To avoid an abrupt jump in July 1970, I multiplied the second series by the factor of 1.43, which is the average ratio of the first to the second series during the overlapping interval from March 1967 through June 1970. Data on the value of all New York Stock Exchange Stocks are end-of-month and come from NYSE publications.

REFERENCES

Campbell, John Y. and Kyle, Albert, "Smart Money, Noise Trading, and Stock Price Behavior," mimeo., Princeton University, 1988.

______ **and Shiller, Robert J.,** "Cointegration and Tests of Present Value Models," *Journal of Political Economy*, October 1987, *95*, 1062–88.

______ **and** ______, "Stock Prices, Earnings, and Expected Dividends," *Journal of Finance*, July 1988, *43*, 661–76.

Cecchetti, Stephen G., "The Case of the Negative Nominal Interest Rates: New Estimates of the Term Structure of Interest Rates During the Great Depression," *Journal of Political Economy*, December 1988, *96*, 111–41.

Christie, Andrew A., "The Stochastic Behavior of Common Stock Variances: Value, Leverage, and Interest Rate Effects," *Journal of Financial Economics*, December 1982, *10*, 407–32.

Cochrane, John H., "How Big Is the Random Walk in GNP?" *Journal of Political Economy*, October 1988, *96*, 893–920.

Cohen, Jacob, "Federal Reserve Margin Requirements and the Stock Market," *Journal of Financial and Quantitative Analysis*, September 1966, *1*, 30–54.

DeLong, Bradford J., Shleifer, Andrei, Summers, Lawrence H. and Waldman, Robert J. "The Economic Consequences of Noise Traders," NBER Working Paper no. 2395, October 1987.

Douglas, George W., "Risk in the Equity Markets: An Appraisal of Market Efficiency," *Yale Economic Essays*, Spring 1969, *9*, 3–45.

Eckardt, Walter L. Jr. and Rogoff, Donald L., "100% Margins Revisited," *Journal of Finance*, June 1976, *31*, 996–1000.

Edwards, Franklin R., "The Crash: A Report on the Reports," in *Information Technology and Securities Markets Under Stress*, Center for Research on Information Systems, NYU Business School Conference, May 1988.

Estrella, Arturo, "Consistent Margin Requirements: Are They Feasible?" *Quarterly Review*, Federal Reserve Bank of New York, Summer 1988, *13*(2), 61–79.

Fama, Eugene F. and French, Kenneth R., (1988a) "Permanent and Temporary Components of Stock Prices," *Journal of Political Economy*, April 1988, *96*, 246–73.

______ and ______, (1988b) "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, October 1988, *22*, 3–25.

Flood, Robert P., Hodrick, Robert J. and Kaplan, Paul, "An Evaluation of Recent Evidence on Stock Market Bubbles," NBER Working Paper no. 1971, July 1986.

French, Kenneth R., Schwert, G. William and Stambaugh, Robert F., "Expected Stock Returns and Volatility, *Journal of Financial Economics*, February 1987, *19*, 3–29.

Friedman, Milton, "The Case for Flexible Exchange Rates," in *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Froot, Kenneth A., "Tests of Excess Forecast Volatility in the Foreign Exchange and Stock Markets," NBER Working Paper no. 2362, August 1987.

Fuller, Wayne A., *The Statistical Analysis of Time Series*, New York: Wiley & Sons, 1976.

Garbade, Kenneth D., "Federal Reserve Margin Requirements: A Regulatory Initiative to Inhibit Speculative Bubbles," in Paul Wachtel, ed., *Crises in Economic and Financial Structure*, Lexington, MA: Lexington Books, 1982.

Grube, R. Corwin, Joy, O. Maurice and Panton, Don B., "Market Responses to Federal Reserve Changes in the Initial Margin Requirement," *Journal of Finance*, June 1979, *34*, 659–74.

Hansen, Lars P., "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, July 1982, *50*, 1029–54.

Hardouvelis, Gikas A., (1988a) "Evidence on Stock Market Speculative Bubbles: Japan, United States, and Great Britain," *Quarterly Review*, Federal Reserve Bank of New York, Summer 1988, *13*(2), 4–16.

______, (1988b) "Margin Requirements and Stock Market Volatility," *Quarterly Review*, Federal Reserve Bank of New York, Summer 1988, *13*(2), 80–89.

______, (1988c) "Margin Requirements, Volatility, and the Transitory Component of Stock Prices," First Boston Working Paper no. 88-38, Columbia University Graduate School of Business, November 1988.

______ **and Steve Peristiani**, "Do Margin Requirements Matter? Evidence from U.S. and Japanese Stock Markets," *Quarterly Review*, Federal Reserve Bank of New York, Winter 1989–90, *14*(4), 16–35.

Harris, Lawrence, "S&P 500 Futures and Cash Stock Price Volatility," mimeo., University of Southern California, School of Business Administration, October 1987.

Hart, Oliver D. and Kreps, David M., "Price Destabilizing Speculation," *Journal of Political Economy*, October 1986, *94*, 927–52.

Largay, James A., "100% Margins: Combating Speculation in Individual Security Issues," *Journal of Finance*, September 1973, *28*, 973–86.

______ **and West, Richard R.,** "Margin Changes and Stock Price Behavior," *Journal of Political Economy*, March/April 1973, *81*, 328–39.

Lehmann, Bruce N., "Fads, Martingales, and Market Efficiency," NBER Working Paper no. 2533, March 1988.

Lo, Andrew W. and McKinlay, Craig A., "Stock Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, Spring 1988, *1*, 41–66.

Luckett, Dudley G., "On the Effectiveness of the Federal Reserve's Margin Requirement," *Journal of Finance*, June 1982, *37*, 783–95.

Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Malkiel, Robert G., "The Capital Formation Problem in the U.S.," *Journal of Finance*, May 1976, *34*, 369–81.

Merton, Robert C., "On Estimating the Expected Return on the Market," *Journal of Financial Economics*, December 1980, *8*, 323–61.

Moore, Thomas G., "Stock Market Margin Requirements," *Journal of Political Economy*, April 1966, *74*, 158–67.

Newey, Whitney K. and West, Kenneth D. "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, May 1987, *55*, 703–08.

O'Brien, James M., "Speculative Bubbles in Stock Prices and the Need for Margin Regulation," mimeo., Board of Governors of the Federal Reserve System, presented at the American Finance Association meetings, December 1984.

______, "Testing for Transient Elements in Stock Prices," mimeo., Board of Governors of the Federal Reserve System, December 1987.

Officer, R. R., "The Variability of the Market Factor of the New York Stock Exchange," *Journal of Business*, July 1973, *46*, 434–53.

Pindyck, Robert S., "Risk, Inflation, and the Stock Market," *American Economic Review*, June 1984, *74*, 335–51.

Poterba, James M. and Summers, Lawrence H., "The Persistence of Volatility and Stock Market Fluctuations," *American Economic Review*, December 1986, *76*, 1142–51.

______ **and** ______, "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics*, October 1988, *22*, 27–59.

Schwert, G. William, Effects of Model Specification on Tests for Unit Roots in Macroeconomic Data," *Journal of Monetary Economics* July 1987, *20*, 73–103.

______, (1988a) "Why Does Stock Market Volatility Change Over Time?" University of Rochester, William Simon Graduate School of Business Working Paper 87-11, May 1988.

______, (1988b) "Business Cycles, Financial Crises and Stock Volatility," University of Rochester, William Simon Graduate School of Business Working Paper 88-06, October 1988.

Scott, Luis O., "The Present Value Model of Stock Prices: Regression Tests and Monte Carlo Results," *Review of Economics and Statistics*, November 1985, *67*, 599–605.

Shiller, Robert J., "Stock Prices and Social Dynamics," *Brookings Papers on Economic Activity*, Fall 1984, *2*, 457–510.

______ **and McCulloch, J. Huston,** "The Term Structure of Interest Rates," NBER Working Paper no. 2341, August 1987.

Sofianos, George, "Margin Requirements on Equity Instruments," *Quarterly Review*, Federal Reserve Bank of New York, Summer 1988, *13(2)*, 47–60.

Summers, Lawrence H., "Does the Stock Market Rationally Reflect Fundamental Values?" *Journal of Finance*, Proceedings, July 1986, *41*, 591–600.

West, Kenneth D., "Dividend Innovations and Stock Price Volatility," *Econometrica*, January 1988, *56*, 37–61.

Federal Reserve System, Board of Governors, *A Review and Evaluation of Margin Requirements*, Staff Study, December 1984.

______, *Annual Report*, various issues.

GAUSS, Applied Technical Systems, Kent, Washington, 1986.

RATS, VAR Econometrics, Inc., Evanston, IL: 1987.

Stocks, Bonds, Bills, and Inflation: 1988 Yearbook, Ibbotson Associates, Chicago, IL, 1988.

Business Cycle Models with Endogenous Technology

By GEORGE W. STADLER *

This paper compares real and monetary business cycle models with and without endogenous technical change. If technology is endogenous, the properties of these models change significantly. In particular, both real and monetary models yield very similar output processes if growth is endogenous, and changes in aggregate demand can result in permanent changes in productivity, employment, and output. The effect of depreciation of technology is examined, and the pattern of real wage movements over the cycle when money wages are fixed but technology is changing is briefly considered. (JEL *131).*

This paper examines the implications that endogenous technical change has for business cycle theory. Conventionally, both real and monetary models of the business cycle assume that the evolution of technology is exogenous to the economic system.[1] The monetary models ignore technical change, thereby implying that the demand-side disturbances that give rise to economic fluctuations in these models have no impact on technology. The real business cycle models assume that random changes in or "shocks" to productivity or technology cause output fluctuations in perfectly competitive environments. These temporary or permanent shocks to productivity are modeled as random drawings from a probability distribution: they are independent of economic forces and conditions and are thus exogenous to the economic environment.

However, there exists a substantial theoretical and empirical literature that emphasizes the dependence of technological progress on economic factors, implying that technology is, to a large degree, endogenous. The work of Kenneth Arrow (1962) points to the importance of learning by doing, and W. A. Eltis (1971) emphasizes the role of expenditure on research and development in generating economic growth. Jacob Schmookler's (1966) pioneering study on the interrelationship between market conditions and innovation led him to conclude that "the belief that invention, or the production of technology generally, is in most instances a noneconomic activity, is false."[2] James Utterback (1974) made a similar finding and reported:

> Market forces appear to be the primary influence on innovation. From 60 to 80 percent of important innovations in a large number of fields have been in response to market demands or needs. The remainder have originated in response to new scientific or technological advances and opportunities. [p. 621]

*Department of Economics, University of Newcastle upon Tyne, NE1 7RU, England. I am indebted to Zvi Hercowitz, David Laidler, Bennett McCallum, Michael Parkin, Ian Wooton, an anonymous referee, and in particular to Peter Howitt for helpful comments and suggestions. Any errors are solely my responsibility.

[1] Typical real business cycle models are those of John Long and Charles Plosser (1983) and Finn Kydland and Edward Prescott (1982). Robert King et al. (1988a) argue that the neoclassical model of capital accumulation (augmented by shocks to productivity) is the basic framework for real business cycle analysis, and discuss its limitations. The monetary models fall into two groups: the misperception models originating with the work of Robert Lucas (1973) and Robert Barro (1976), and models that emphasize nominal rigidities, such as Stanley Fischer (1977) or John Taylor (1980).

[2] Schmookler (1966, p. 208). Other studies have also found that economic factors have a significant influence on technology—Morton Kamien and Nancy Schwarz (1982) provide a survey of the literature.

Thus, exogenous changes in the scientific base (the "shocks" of real business cycle theory) do not appear to be the cause of most of the changes in productive techniques.

Conventional real and monetary business cycle theory (with exogenous technology) has ignored this body of literature. It is therefore interesting to see how real and monetary business cycle models with *endogenous* technology differ from the conventional models, and to what extent the results obtained from conventional models are sensitive to the assumption that technical change is independent of economic conditions.[3]

There is a second motivation for examining models of business cycles with endogenous technology. Charles Nelson and Charles Plosser (1982) first observed that most economic time-series appear to be nonstationary and are best modeled as containing stochastic trends. This implies that some fraction of innovations in output is permanent and alters the long-run trend of output. Traditionally, monetary innovations were considered to have only a transitory impact on output: the influence of a monetary shock dissipates over time, leaving the trend of output unchanged. Thus, models based on monetary innovations (or demand-side innovations generally) have difficulty in accounting for nonstationarity. This has fostered the growing belief that monetary models of the business cycle are unable to provide a satisfactory explanation of output fluctuations and should be rejected in favor of real business cycle models.[4]

This paper shows that this inference is false. If technology is modeled as endogenous, a monetary business cycle model exhibits properties that are very different from those of the conventional models of Robert Lucas and Robert Barro or John Taylor and Stanley Fischer. There is a long-run nonneutrality of money in models with endogenous technology: monetary shocks have a permanent impact on output, and output is nonstationary, even in the absence of exogenous shocks to the supply side. The reason is that changes in demand that raise the level of output can, through a number of channels, exert a permanent influence on the supply side. For example, changes in the utilization of factor inputs when demand changes can result in reorganization and the acquisition of new skills; or a higher level of output may make innovation more profitable and result in the allocation of more resources to R&D. Furthermore, even real business cycle models with endogenous technology have somewhat different properties than conventional real business cycle models.

The following section develops a general model that contains both real and monetary disturbances. Unanticipated changes in the money supply have real effects in this model economy because nominal wages are as-

[3] A number of recent papers recognize the importance of endogenous technology in accounting for economic growth, for example, Lucas (1988), Edward Prescott and John Boyd (1987), and Paul Romer (1986), but these papers do not examine endogenous technology in a cyclical context. However, Robert King and Sergio Rebelo (1986) examine the impact of productivity shocks in a real business cycle model with endogenous growth, and their key conclusions are similar to those implied by the model of Section II, Part C below. This paper differs from King and Rebelo in that, first, they model the growth process as resulting from investment in human and physical capital, while I assume that changes in technological knowledge depend on the level of economic activity, specifically the level of labor input; second, this paper allows a role for money and aggregate demand shocks; third, it explicitly compares real and monetary business cycle models with and without endogenous growth. The model and conclusions of King and Rebelo's paper are briefly outlined in Section 3 of King et al. (1988b).

[4] For example, see Nelson and Plosser (1982, p. 159). Even Barro (1984) has suggested that the lasting contribution of the rational expectations revolution seems to lie in the development of the real, not monetary, business cycle theory. Bennett McCallum (1986), however, assesses the evidence against monetary models and concludes that it is insufficient to justify a rejection of this class of models. Furthermore, Kenneth West (1988) has shown that a purely monetary model with overlapping wage contracts can, for certain parameter values, yield an output process that has a root that is below but near unity. This implies that it may be impossible to discriminate in finite samples between a stationary monetary model and a nonstationary real business cycle model. By contrast, this paper demonstrates that with endogenous growth a purely monetary model can give rise to a nonstationary output process with a unit or greater-than-unit root.

sumed to be set in advance by one-period contracts. This is not a crucial assumption. Very similar results would be obtained if we assumed that prices, rather than wages, are sticky, or if a Lucas-type misperception framework had been adopted. Technology is partly endogenous because it is assumed that technical knowledge advances through learning by doing and depends on the level of labor input and on the level of labor productivity achieved in previous periods.

Within the general model are nested various sub-models that enable us to compare real and monetary models with and without endogenous technology. In Section II attention is focused on four sub-models; a pure monetary model with endogenous technology, a typical real business cycle model with exogenous technology, a real business cycle model with endogenous technology, and finally a hybrid model with transitory real and monetary shocks but no learning curve. The properties of the models are contrasted, and the long-run properties of the general model are examined.

The effects of depreciation of technology are examined in Section III. A monetary model with endogenous technology and a real business cycle model with exogenous technology are considered, where both the endogenous and exogenous technologies are subject to depreciation. The properties of a conventional real business cycle model change considerably if technology is subject to depreciation.

The fourth section considers the implications of learning for the pattern of real wages over the business cycle, when money wages are fixed by contract. Normally, a fixed money wage results in countercyclical movements in real wages, since prices move procyclically. However, once learning is introduced, the resulting pattern of real wages over the cycle becomes unclear. Even with nominal contracts, real wages may exhibit very little countercyclical movement.

The final section contains a summary and conclusions.

I. A General Model with Endogenous Technology

This section develops a simple model of output fluctuations with endogenous technology. Following Arrow, the concept of technical change I adopt comes from the basic premise that learning is the product of experience and takes place during activity.

I consider a simple closed economy. The supply side of the commodity market consists of a large number of competitive firms, all producing an identical good as output, that take the market price as given. Trading arrangements are sequential: at the beginning of each period, before the current period's price level is known, households enter into one-period nominal wage contracts with firms, agreeing to provide whatever amount of labor the firm requires at that wage rate. Thus, in order to negotiate the wage contract, agents must form expectations of both the price level and the representative firm's demand for labor before they know the value of the money supply (and aggregate demand) and before the real shocks hitting the economy are observed.

A. *Aggregate Supply*

The representative firm maximizes its stream of discounted expected profits:

$$V = \max_{\{L_t^i\}} E_t\left[\sum_{j=0}^{\infty} \beta^j \left(Y_{t+j}^i - \frac{W_{t+j}}{P_{t+j}} L_{t+j}^i\right)\right] \qquad 0<\beta<1$$

where L_t^i is employment of labor by the representative firm, Y_t^i is the output of the representative firm, W_t is the money wage, P is the price level, and β is a discount factor. The production technology is assumed to be Cobb-Douglas:

$$(1)\quad Y_t^i = k\left(L_t^i\right)^{\alpha} Z_t^{(1-\alpha)} F_t \qquad 0<\alpha<1$$

where k is a constant and F_t is a strictly positive stochastic shock that contains two components: (i) a permanent component, ξ_t, which can be thought of as a shock to technology, and (ii) a temporary productivity shock, η_t, so that in natural logarithms the evolution of F_t is given by

$$(2)\qquad f_t = \bar{f}_t + \eta_t$$

$$\bar{f}_t = \bar{f}_{t-1} + \xi_t,$$

where lowercase letters denote natural logarithms. Both η_t and ξ_t are stationary stochastic processes with zero mean, constant variance, and zero covariance.

Z_t is a scale factor that represents accumulated technical knowledge. It evolves according to

$$(3)\quad Z_t = Z_{t-1}\left[\frac{Y_{t-1}}{L_{t-1}}\right]^{\lambda}(L_{t-1})^{\gamma} \qquad 0<\lambda,\gamma<1.$$

The evolution of technical knowledge depends on the level of aggregate labor input (L_{t-1}) employed the previous period, and on the level of aggregate labor productivity attained the previous period.[5] The greater the level of labor input, the greater is the scope for learning and acquisition of new skills. A higher level of labor input also requires more intensive use of factors fixed in the short run, thus raising the incentive to eliminate waste and bottlenecks. However, some change in productivity can occur independently of changes in labor input, for example, through reorganization of the production process and firm structure to achieve greater efficiency. The benefits of such changes are likely to persist for some time while the improved structures are in place, and to capture this source of productive improvements, I assume that technical knowledge also depends on the level of labor productivity attained in the previous period. For simplicity, it is assumed that Z is disembodied knowledge that does not depreciate (i.e., knowledge in books that is easily acquired by new entrants to the labor market). I abstract from physical capital (although Z is a kind of capital).

Z depends on aggregate labor input and productivity. Since the individual firm is small, I assume it ignores the effect of its own employment decision and level of productivity on aggregate employment and productivity. The learning process can be thought of as a by-product of the output process. It is not necessarily under the conscious control of management. Furthermore, as labor moves between firms, learning skills become dispersed, so that aggregate labor input and productivity determine knowledge.

Assuming that the evolution of knowledge is exogenous to the firm enables us to treat the firm's problem like a one-period problem,[6] and the firm's only decision variable is L_t^i. Taking the first-order condition for labor enables us to derive the representative firm's demand for labor (in logarithms):

$$(4)\quad l_t^{i,d} = \frac{\ln(\alpha k)}{1-\alpha} + \frac{1}{1-\alpha}(p_t - w_t) + z_t + \frac{1}{1-\alpha}f_t.$$

The labor supply function is assumed to be

$$(5)\quad L_t^s = e^{\phi_1}(W_t/P_t)^{\phi_2} \qquad 0<\phi_2<1.$$

At the beginning of the period the money wage is set at its expected market clearing level, that is, the wage at which the expected demand for labor during the period ahead equals the amount households expect to supply. The amount of labor households expect to supply will depend on the ex-

[5] An alternative but more complex way of introducing endogenous technology is to assume the firm faces an innovation production function and that innovation comes about through the application of resources to R&D. The results derived from such a model are similar to those presented below (see George Stadler, 1988). An alternative and simpler way of modeling learning is to assume that technical knowledge just depends on past levels of output, as in Stadler (1986).

[6] Concavity of the production function (both labor and knowledge have diminishing marginal products) and the treatment of learning as an externality ensure that a competitive equilibrium exists each period with firms treating the state of technical knowledge as exogenously given. If learning were internalized, a competitive equilibrium might not exist. Since there is implicitly no limit to learning, utilization of labor yields increasing returns *over time* and the growth rate of output will rise over time. The assumptions that underpin this result are examined in Section II, Part E.

pected real wage. Taking the expectation of equation (5) yields, in logarithms,

$$(6)\qquad E_{t-1}l_t^s=\phi_1+\phi_2(w_t-p_t^e),$$

where $p_t^e\equiv E_{t-1}p_t$ is the rational expectation of the price level. The rational expectation of the firm's demand for labor formed at the beginning of the period is obtained by taking the expectation of equation (4), and using equations (2) and (3) to obtain

$$(7)\quad E_{t-1}l_t^{i,d}=\frac{\ln(\alpha k)}{1-\alpha}+z_t+\frac{1}{1-\alpha}(p_t^e-w_t)+\frac{1}{1-\alpha}\bar{f}_{t-1},$$

since $E_{t-1}z_t=z_t$ and $E_{t-1}f_t=\bar{f}_{t-1}$, because Y_{t-1}, L_{t-1}, Z_{t-1}, and $\bar{f}_{t-1}$ are contained in the information set available at the beginning of period t, and $E_{t-1}\eta_t=E_{t-1}\xi_t=0$. Equating the expected supply and expected demand for labor given by the above two equations yields the money wage written into the labor contract:

$$(8)\quad w_t=p_t^e+\frac{\ln(\alpha k)-(1-\alpha)\phi_1}{1+(1-\alpha)\phi_2}+\frac{(1-\alpha)z_t}{1+(1-\alpha)\phi_2}+\frac{\bar{f}_{t-1}}{1+(1-\alpha)\phi_2}.$$

Substituting equation (8) into equation (4) for w_t gives the firm's employment of labor:

$$(9)\qquad l_t^i=a_0+a_1(p_t-p_t^e)+a_2z_t+a_1f_t-a_3\bar{f}_{t-1}$$

$$a_o=\frac{\ln(\alpha k)\phi_2+\phi_1}{1+(1-\alpha)\phi_2}$$

$$a_1=\frac{1}{1-\alpha}$$

$$a_2=1-\frac{1}{1+(1-\alpha)\phi_2}$$

$$a_3=\frac{1}{(1-\alpha)+(1-\alpha)^2\phi_2}.$$

Substituting (9) into a logarithmic form of the production function given by equation (1) and aggregating gives us the aggregate supply function:

$$(10)\quad y_t^s=q+b_0+b_1(p_t-p_t^e)+b_2z_t+b_3(\xi_t+\eta_t)+b_4\bar{f}_{t-1}$$

$$b_0=\ln(k)+\alpha a_0$$

$$b_1=\alpha/(1-\alpha)$$

$$b_2=(1-\alpha)+\alpha a_2$$

$$b_3=1/(1-\alpha)$$

$$b_4=\frac{(1+\phi_2)}{1+(1-\alpha)\phi_2},$$

where q is the natural logarithm of the number of firms in the economy. Hence aggregate supply will increase in response to a positive price surprise, a positive real shock, or a rise in technical knowledge.

B. *Aggregate Demand and Expectations*

I assume that aggregate demand is given by the quantity theory equation with velocity constrained to unity:

$$(11)\qquad Y_t^d=M_t/P_t,$$

where Y_t^d is real aggregate demand and M_t is the money stock.

The money supply is assumed to follow a random walk with positive trend. The trend is set by the central bank so as to avoid an ongoing deflation in a growing economy, given that the velocity of money is constant. The money supply rule is public knowledge and, in logarithms, is

$$(12)\qquad m_t=m_{t-1}+\mu+\varepsilon_t,$$

where μ is a constant trend and ε_t is a zero-mean stochastic error with constant variance.

When signing wage contracts, agents have at their disposal the past values of all relevant variables (which enables them to calculate z_t), including $\bar{f}_{t-1}$. Writing (11) in natural logarithms, rearranging terms, and taking expectations gives us the rational expectation of the price level:

$$(13)\qquad p_t^e = m_t^e - y_t^{de},$$

where $y_t^{de} \equiv E_{t-1} y_t^d$ is the rational expectation of aggregate demand formed at the beginning of the period. Equations (11) and (13) imply

$$(14)\quad (p_t - p_t^e) = (m_t - m_t^e) - (y_t^d - y_t^{de}).$$

Expected market clearing implies that $y_t^{de} = y_t^{se}$, so that one can take the expectation of equation (10) to obtain y_t^{de}:

$$(15)\quad y_t^{de} = y_t^{se} = q + b_0 + b_2 z_t + b_4 \bar{f}_{t-1},$$

since $E_{t-1}(p_t - p_t^e) = 0$, $\bar{f}_{t-1}$ is contained in the information set available at the beginning of the period, and the expected values of the real shocks (η_t and ξ_t) are zero. Actual market clearing implies $y_t^d = y_t^s$, so that subtracting equation (15) from equation (10) gives an expression for $(y_t^d - y_t^{de})$, the difference between actual and expected aggregate demand:

$$y_t^d - y_t^{de} = b_1(p_t - p_t^e) + b_3(\eta_t + \xi_t).$$

Substituting this expression into equation (14) and noting that $m_t - m_t^e = \varepsilon_t$ yields an expression for the difference between the actual and expected price level:

$$(16)\quad (p_t - p_t^e) = \frac{1}{1+b_1}[\varepsilon_t - b_3(\eta_t + \xi_t)].$$

Thus prices diverge from their expected values each period because of unanticipated real and monetary shocks.

C. *The Output Process*

The level of output each period is obtained by substituting equation (16), which gives the price surprise into equation (10), the aggregate supply function:

$$(17)\quad y_t = q + b_0 + \alpha[\varepsilon_t - b_3(\eta_t + \xi_t)] + b_2 z_t + b_3(\eta_t + \xi_t) + b_4 \bar{f}_{t-1}.$$

Output depends on technology, both on the level of technical knowledge, z_t, and on the exogenous shocks to technology. However, real shocks are partly offset through price level effects. A positive value of η_t or ξ_t raises output and requires the price level to fall below its expected value to clear the commodity and money markets. The fall in the price level raises the real wage, thus counteracting the effect of the productivity shocks on employment and on output to some extent. Output also depends on unanticipated changes in the money supply, ε_t, and, as in all demand-side models of the business cycle with nominal wage contracting, an unanticipated rise in the money supply causes the price level to rise above its expected value and hence depresses the real wage and raises output.

The output process of equation (17) is nonstationary. The nonstationarity arises from two sources. First, as can be seen from equation (2), $\bar{f}_{t-1}$ is nonstationary because it consists of accumulations of permanent stochastic shocks. Second, z_t is nonstationary for, as can be seen by solving equation (3) backward, it depends on the cumulative sum of past levels of output and employment. The properties that each of these two sources of nonstationarity imposes on the output process become clearer in the next section, where a number of sub-models, containing some but not all the properties of the general model, are considered.

II. A Comparison of Real and Monetary Models with Endogenous and Exogenous Technology

Nested within the general model of equation (17) are at least four sub-models: a purely monetary model that contains no real shocks but with endogenous technology; real models with and without endogenous technology; and a hybrid model with monetary shocks and transitory real shocks but no learning. Each of these models is considered in turn, beginning with the monetary model with endogenous technology.

A. *A Monetary Model with Endogenous Technology*

This subsection considers a purely monetary model that contains no real shocks. The only disturbances to the economic system are unanticipated changes in the money supply, so that this model is a special case of the general model with $F_t = 1$ for all t (i.e., $\eta_t = \xi_t = 0$ for all t). The difference between actual and expected prices now only depends on innovations in the money supply:

$$(p_t - p_t^e) = \frac{1}{1+b_1}\varepsilon_t,$$

and the output path is given by

$$y_t = q + b_0 + \alpha\varepsilon_t + b_2 z_t. \tag{18}$$

This output process is nonstationary and has a greater-than-unit root. This can be seen more clearly if one takes the first difference of output, uses equation (3) and an aggregate version of equation (9), and, after some manipulation, writes y_t as a moving average process:

$$y_t = (1+\tau)^t y_0 + \alpha\varepsilon_t + \sum_{j=0}^{t-1}(1+\tau)^j c_0 + c_1\sum_{j=0}^{t-2}(1+\tau)^j\varepsilon_{t-j-1} \tag{19}$$

$$\tau = \left[1 - \frac{1}{1+(1-\alpha)\phi_2}\right]\gamma + \left[\frac{1-\alpha}{1+(1-\alpha)\phi_2}\right]\lambda, \quad \tau > 0;$$

$$c_0 = (b_2 - a_2)(\gamma-\lambda)q + b_2(\gamma-\lambda)a_0 - a_2(\gamma-\lambda)b_0;$$

$$c_1 = [\tau\alpha + (1-\alpha)(\gamma-\lambda)], \quad c_1 > 0,$$

where y_0 is a reference value in the past and I assume $\varepsilon_0 = 0$ for simplicity.

Equation (19) demonstrates that output depends on deterministic components of growth, captured by the constant, c_0, and the term $(1+\tau)^t y_0$. With endogenous technology the economy will grow even in the absence of stochastic shocks, because utilization of labor results in an ongoing learning process that increases technical knowledge, leading to higher productivity and output in future periods.

Output also depends on a nonstationary, stochastic growth component that consists of the cumulative sum of money-supply shocks that cause errors in price expectations. However, these errors have a permanent effect on output. Initially, a positive innovation in the money stock raises output by $\alpha\varepsilon_t$. In the absence of further shocks, output falls the following period, but not back to its previous level, for through learning a monetary innovation of ε_t raises output by $c_1\varepsilon_t$ the next period. Thereafter, the impact of the shock gradually increases over time; in j periods ahead of it, output will have increased by $(1+\tau)^{j-1}c_1\varepsilon_t$. Thus, innovations in output are, to some degree, persistent, but the degree of persistence depends on the horizon from which the innovation is viewed: with endogenous technology, the impact of an innovation in output slowly rises over time.[7] The reason for this is that any shock that causes an innovation in output adds a stimulus to an ongoing, endogenous growth process and raises the time profile of output permanently above

[7] Empirical studies have found differing estimates of persistence of innovations in output. John Campbell and Gregory Mankiw (1987) find that an unanticipated change in real GNP of 1 percent should change one's long-run forecast of output by *more* than 1 percent—between 1.3 and 1.9 percent. Other studies, such as John Cochrane (1988) and Mark Watson (1986), have found less persistence. Cochrane finds that at most 0.4 percent of an innovation in GNP is permanent, and Mark Watson finds 0.6 percent using an unobserved components model, but considerably more (1.7 percent) using an ARIMA model.

the level it would have achieved in the absence of the shock. A monetary shock at time t that raises employment and output above their steady-state values will, through learning, cause technical knowledge to rise above its steady-state value by time $t+1$. This rise in technology will increase the productivity of and demand for labor above their steady-state values at time $t+1$, bidding up the expected real wage and level of employment. Hence, as a result of a monetary innovation at time t, through learning, output will be above its steady-state value at time $t+1$ as well. However, the rise in labor input and output at time $t+1$ causes a *further* increase in technical knowledge above its steady-state value by time $t+2$. This results in a further rise in labor productivity and employment at time $t+2$, enabling further acquisition of technical skills by time $t+3$, which raises productivity and employment in that period, and so on. Thus, the endogenous growth process not only acts as a propagation mechanism for aggregate demand shocks but also amplifies their impact over time.

In contrast to conventional business cycle models, the sub-model captured by equation (19), like the general model it is derived from, has a greater-than-unit root in the output process (of value $(1+\tau)$) and is not difference stationary. Since (19) is in logarithmic form, the greater-than-unit root implies that in natural levels the growth rate of output will increase over time. This arises because, as discussed in Section II, Part E below, the growth rate of output essentially depends on the level of output.

Furthermore, the natural rate of unemployment is likely to depend on the history of aggregate demand in this model. The "natural" level of employment at any point in time can be regarded as the level of employment that prevails when there is no discrepancy between labor's actual and expected real wage. A sufficient condition for this is that there be no real or monetary surprises that drive a wedge between the actual and expected price level; that is, $\varepsilon_t = \xi_t = \eta_t = 0$. Aggregating equation (9) and using (3) to solve backward for z_t, we can obtain the following expression for l_t:

$$(20)\quad l_t = q + a_0 + a_2 z_0 + a_2\lambda \sum_{j=1}^{t} y_{t-j} + a_2(\gamma-\lambda)\sum_{j=1}^{t} l_{t-j},$$

where z_0 is a reference value in the past. Thus, the natural level of employment at time t is a stochastic variable; it depends on the cumulative sum of past values of employment and on the cumulative sum of past values of output, which, from equation (19), depends on the cumulative sum of innovations in aggregate demand (and, in the more general model, on exogenous real shocks as well). The demand for labor and the equilibrium level of employment are consequently not invariant with respect to the history of aggregate demand in the economy. The natural rate of unemployment, being the difference between the supply of labor and the equilibrium level of employment, is likely to change in response to permanent changes in employment and the demand for labor.[8] This suggests that in an economy with endogenous technology, changes in aggregate demand are likely to alter the natural rate of unemployment.

The results obtained from this model stand in strong contrast to the results derived from conventional monetary business cycle models. In both the misperception models and in the wage-contracting models, money has a transitory influence on eco-

[8] A general rise in the demand for labor resulting from a higher level of technical knowledge can reduce the natural rate of unemployment if it induces firms to establish training programs in areas of skill shortages. This could reduce the mismatch of skills required by employers and those possessed by the unemployed, hence lowering structural unemployment. A rise in demand for labor can also lead to greater expenditure by firms on recruitment and advertising, so leading to improved information flows and possibly lower search unemployment.

nomic activity: the impact of a monetary innovation dies away over time, and, in the long run, money is neutral. However, when technology is dependent on demand conditions, monetary shocks have a permanent impact on output. There is a long-run nonneutrality of money in models with endogenous technology.

B. *A Real Business Cycle Model with Exogenous Technology*

A second sub-model nested within the general model of equation (17) is a pure real business cycle model with technology evolving exogenously. Money has no impact on economic activity, and the model can be derived from the general model of equation (17) by assuming that not only the velocity but also the quantity of money is constant, so that the level of nominal aggregate demand is constant. In this case, unexpected price level changes are caused by technology shocks that alter real output. As is well known, the price level is countercyclical in such a model.

In the absence of endogenous technology, without a learning curve, $Z_t = 1$ for all t. Proceeding as before, the price surprise term reduces to

$$(21)\quad (p_t - p_t^e) = -[b_3/(1+b_1)](\eta_t + \xi_t),$$

for by assumption $\varepsilon_t = 0$ for all t, and the value of output is

$$(22)\quad y_t = q + b_0 - \alpha b_3(\eta_t + \xi_t) + b_3(\eta_t + \xi_t) + b_4 \bar{f}_{t-1}.$$

Output depends on exogenous real shocks. As in the general model, a positive innovation in η_t or ξ_t has a positive effect on output. The positive output effect is offset to some extent by the effect of the fall in the price level on the real wage. Taking the first difference of equation (22) and using (2) to solve backward yields

$$(23)\quad y_t = y_0 + \eta_t + \xi_t + b_4 \sum_{j=1}^{t-1} \xi_{t-j},$$

where y_0 is a reference value at time $t = 0$ and I have assumed that $\eta_0 = \xi_0 = 0$.

Output is clearly nonstationary in this model, for it depends on the cumulative sum of permanent shocks to technology, the ξ's. As in other real business cycle models, exogenous changes in technology cause permanent shifts in output, for the influence of the ξ's does not decay over time. The initial impact of a positive innovation in ξ is, from equation (22), to raise output by $b_3(1-\alpha)\xi_t$. In the following and all future periods, it raises output by $b_4\xi_t$, so that its long-run impact exceeds the short-run impact ($b_4 > (1-\alpha)b_3$), for in the long run the offsetting rise in the real wage is absent.

C. *A Real Business Cycle Model with Endogenous Technology*

A third model of interest is a pure real business cycle model with endogenous technology, that is, a model in which learning occurs but money has no real effects. The production function of the representative firm is given by equation (1), and derivation of the model proceeds as in the case of the general model above, but with $\varepsilon_t = 0$ for all t, so that again the level of nominal aggregate demand is held constant. Thus, the only difference between this model and the preceding one is that technical knowledge, Z_t, rises over time in response to labor input and increases in productivity.

In the absence of monetary shocks, the difference between actual and expected prices depends only on real shocks and is given by equation (21). The output equation is very similar to that of the general model:

$$(24)\quad y_t = q + b_0 - \alpha b_3(\xi_t + \eta_t) + b_2 z_t + b_3(\xi_t + \eta_t) + b_4 \bar{f}_{t-1}.$$

Using equations (3) and (9) enables us to write, after some manipulation,

$$
\begin{aligned}
(25)\quad y_t = {} & (1+\tau)^t y_0 \\
& + \sum_{j=0}^{t-1} (1+\tau)^j c_0 + \eta_t + \xi_t \\
& + c_2 \sum_{j=0}^{t-2} (1+\tau)^j \xi_{t-j-1} \\
& + c_3 \sum_{j=0}^{t-2} (1+\tau)^j \eta_{t-j-1} \\
& + c_4 \sum_{j=0}^{t-1} (1+\tau)^j \bar{f}_0 \\
& + c_4 \sum_{i=0}^{t-2} \sum_{j=i}^{t-2} (1+\tau)^j \xi_{t-j-1}
\end{aligned}
$$

$$c_2 = b_2\lambda + b_4 + [b_2(a_3-a_1) + a_2 b_4](\gamma-\lambda);$$

$$c_3 = b_2\lambda; \quad c_4 = [b_2(a_1-a_3) - a_2 b_4](\gamma-\lambda),$$

where $\bar{f}_0$ is a reference value in the past and I assume $\eta_0 = \xi_0 = 0$.

Output depends on the cumulative sums of the exogenous shocks and is nonstationary. It thus bears some resemblence to the conventional real business cycle model of equation (23). Nevertheless, there are two important differences between this model and the real business cycle model with exogenous technology. First, like the pure monetary model (and like the general model), this model has a greater-than-unit root, the value of the root being $(1+\tau)$. The conventional model of equation (23) has a *unit* root, and is difference stationary, while the model with endogenous technology is not difference stationary. As in the monetary model, in a real business cycle model with endogenous technology the growth rate of output increases over time.

Second, with endogenous technology, transitory shocks to productivity (the η's) have a permanent effect on output. Again, learning acts as a propagation mechanism for both demand-side and supply-side disturbances to output. A temporary rise in productivity stimulates learning and causes a rise in Z, shifting the long-run trend of output upward. Thus, in a model with endogenous growth, output will have a stochastic trend even if the only disturbances hitting the economy are transitory changes in productivity. (This is also the main finding of Robert King and Sergio Rebelo, 1986.)

D. *A Hybrid Model*

A final model that deserves consideration is one that contains both monetary shocks and transitory real shocks to technology but has no learning curve. Hence, $Z_t = 1$ for all t, and $F_t = e^{\eta_t}$, so that the firm's production function becomes

$$Y_t^i = k(L_t^i)^{\alpha} e^{\eta_t},$$

and the output process is governed by

$$(26)\quad y_t = q + b_0 + \alpha[\varepsilon_t - b_3\eta_t] + b_3\eta_t.$$

The trend, or expected future value of output conditional on current information, is

$$E_t y_{t+i} = q + b_0; \quad i > 1,$$

since $E_t\varepsilon_{t+i} = E_t\eta_{t+i} = 0$, for these are random variables with zero mean. It is at once apparent that this process is stationary, for output will only vary about the point $(q + b_0)$. This is true even if the monetary and real shocks have effects that persist for several periods before dying away, through, for example, the presence of adjustment costs. In the long run, money is neutral in this model. Thus, in a model with exogenous technology neither monetary shocks nor transitory shocks to technology have a permanent impact on output, thus highlighting the distinctions between models with endogenous and models with exogenous technology.

E. *The Steady State*

In the steady state neither real nor monetary shocks impinge on the economy—$\varepsilon_t = \eta_t = \xi_t = 0$ and $p_t = p_t^e$ for all t. The output process for the general model can then be written as a first-order difference equation:

$$y_t = (1+\tau)y_{t-1} + c_0. \tag{27}$$

The solution to this difference equation is

$$y_{t+n} = [y_t + (c_0/\tau)](1+\tau)^n - (c_0/\tau).$$

This equation describes an explosive growth process, for $\tau > 0$. The growth rate of the natural level of output, viewed n periods ahead at time $t+n$, will be $(1+\tau)^n[y_t + (c_0/\tau)]\ln(1+\tau)$, which increases as n rises. The rising growth rate occurs because the growth rate of technology, as expressed by equation (3), depends on the level of labor input and labor productivity attained. Hence the *growth rate* of technology essentially depends on its *level*, since the amount of labor utilized and labor productivity both depend on Z. This gives rise to the greater-than-unit roots that these models with learning contain; as the level of technology rises, so do productivity and labor input, causing the growth rate of technology to increase. The rise in Z results in a rising marginal product of labor that bids up the real wage and results in a rising level of employment. This can be seen from the long-run time paths of these variables:

$$z_t = (1+\tau)z_{t-1} + c_z$$

$$\ln(\text{MPL}_t) = \ln(\alpha) + y_t - l_t$$

and

$$\Delta\ln(\text{MPL}_t) = (b_2 - a_2)[\tau z_{t-1} + c_z]$$

$$l_t - l_{t-1} = a_2(\tau z_{t-1} + c_z)$$

$$\Delta(w_t - p_t) = c_w(\tau z_{t-1} + c_z),$$

where $c_z = \lambda(q + b_0) + (\gamma - \lambda)a_0 > 0$; $c_w = (1-\alpha)/(1+(1-\alpha)\phi_2) > 0$; $(b_2 - a_2) > 0$. Thus the model exhibits rising growth rates of output, technical knowledge, productivity, employment, and real wages, since the growth rates of these variables depend on the growth rate of technology and thus on the level of technology.

These properties of the model may hold for considerable time periods and can account for the rising growth rates that Paul Romer (1986) reports for the past two centuries. However, these growth patterns may change substantially in the very long run, because they are based on the short-run labor supply function given by equation (5), and (5) is unlikely to hold in the very long run.

As society becomes wealthier, the income effect of a rise in wages may increasingly come to dominate the substitution effect, resulting in little or no increase in work effort as wages rise. Indeed, the long-run supply of labor schedule is probably backward-bending and the long-run value of ϕ_2, the wage elasticity of labor supply, probably negative.[9] However, the value of the root of output depends on the labor supply elasticity ϕ_2, and τ falls as ϕ_2 falls.[10] As $\phi_2 \to -\lambda/\gamma$, $\tau \to 0$, and if ϕ_2 takes the value $-\lambda/\gamma$, $\tau = 0$ and the output process changes to a nonexplosive unit-root process. In this case the time paths of the variables

[9] Daniel Hamermesh and Albert Rees conclude their discussion of labor supply by noting that "The long-run supply curve of employee hours to the economy as a whole...will probably be backward-sloping at the average real wage levels of developed countries.... The backward slope arises from the tendency in developed countries for hours actually worked per year to be reduced as real income rises over time while the overall rate of labor-force participation for all groups in the population taken together remains roughly constant or rises slightly." [1984, p. 82]

[10] $\tau = [1 - (1/(1+(1-\alpha)\phi_2))]\gamma + [(1-\alpha)/(1+(1-\alpha)\phi_2)]\lambda$, and $\partial\tau/\partial\phi_2 > 0$ if $\gamma > (1-\alpha)\lambda$. This condition must hold if a rise in labor input is not, *ceteris paribus*, to have a negative or zero effect on technical knowledge, and is assumed to hold throughout. This can be seen by substituting an aggregate form of equation (1) lagged one period into (3) to obtain $Z_t/Z_{t-1} = k^\lambda Z_{t-1}^{(1-\alpha)\lambda} L_{t-1}^{\gamma-(1-\alpha)\lambda} F_{t-1}^\lambda$. If $\gamma < (1-\alpha)\lambda$, then $\partial(Z_t/Z_{t-1})/\partial L_{t-1} < 0$.

become

$$y_t = y_{t-1} + c_0$$

$$z_t = z_{t-1} + c_z$$

$$\Delta \ln(\mathrm{MPL}_t) = (b_2 - a_2)c_z$$

$$l_t - l_{t-1} = a_2 c_z$$

$$\Delta(w_t - p_t) = c_w c_z$$

and $a_2 < 0$ but $c_z, (b_2 - a_2), c_w > 0$ if $\phi_2 = -\lambda/\gamma$. In this case rising productivity is sufficiently offset by a falling supply of labor to dampen explosive growth.

Thus, if the wage elasticity is sufficiently negative, the model displays long-run features one might expect to observe in a developed economy with roughly constant population: declining employment, manifesting itself as a decline in average hours worked per person employed, steadily rising real wages and productivity, and technological advancement. If the wage elasticity rises above $-\lambda/\gamma$, then $\tau > 0$ and rising growth rates would manifest themselves again, though in conjunction with a decline in average hours worked per person employed if ϕ_2 is still negative. Hence, the actual growth path will depend crucially on the long-run dynamics of labor supply.[11]

A further factor that can dampen the accelerating growth is limits to learning. Continuous learning can only occur if there are exogenous changes to the environment in which agents work. In the absence of new stimuli to learning, growth would slowly peter out. In terms of equation (3), γ and λ would approach zero over time: if $\gamma = \lambda = 0$, Z is constant and $\tau = c_0 = c_z = 0$, so growth ceases and the economy enters the stationary state.

Two pieces of evidence lend some support to my formulation of the learning process and the output process that results from it. First, Romer (1986) presents empirical evidence for several industrial countries that *growth rates* of output have a positive trend and have been rising over the past two centuries. Second, in their survey of technical progress, Charles Kennedy and A. P. Thirlwall (1972) observe that "since product types are continually changing it is probably safe to assume that in aggregate there is no limit to the learning process."[12]

III. Depreciation of Technology

This section examines the properties of business cycle models if technology depreciates. Generally, technical and scientific knowledge is not regarded as subject to depreciation, except under abnormal circumstances such as times of war. However, one can think of certain technologies that, through changes in the state of nature, become obsolete before replacements are found for them. For example, in 1987, *The Economist* reported that "Bacteria are evolving resistance to antibiotics and insects resistance to pesticides faster than new chemicals are being invented. Before long, at this rate, medicine will revert to the days before penicillin and agriculture to the days before DDT."[13] To examine the impact of depreciation in a model with endogenous technology, I allow Z to depreciate over time and replace (3) by

$$(28)\quad Z_t = Z_{t-1}^{(1-\delta)} \left[\frac{Y_{t-1}}{L_{t-1}}\right]^{\lambda} (L_{t-1})^{\gamma} \qquad 0 < \delta < 1$$

where δ is the rate of depreciation of technology.

[11] Changes in population growth will also affect the growth path of output, and a falling labor supply caused by declining population could also dampen accelerating growth. In this case, ϕ_1 in equation (5) would fall over time, causing the constants a_0, b_0, c_0, and c_z to become time-varying.

[12] Kennedy and Thirlwall (1972, p. 39).

[13] March 21, 1987, p. 93. This prognosis was perhaps too pessimistic. In a more recent issue (April 16, 1988) *The Economist* suggests that, at least in agriculture, genetic engineering may overcome some of these problems, implying that the decline in agricultural technology is not permanent.

This section considers two models. The purely monetary model of Section II provides an example of the effect of depreciation in a model with endogenous technology. Second, a real business cycle model with depreciating exogenous technology is appraised. It emerges that such a real business cycle model cannot account for nonstationarity if technology is subject to depreciation.

A. *A Monetary Model with Depreciation*

The path of output in a pure monetary model is given by equation (18). Using equation (28) instead of equation (3) as the process that determines technical knowledge, one can write equation (18) as an ARMA process. Taking the first difference of equation (18) and substituting for l_{t-1}, yields, after some manipulation,

$$(29)\quad y_t = d_0 + \alpha\varepsilon_t - d_1\varepsilon_{t-1} + (1+\tau-\delta)y_{t-1}$$

$$d_0 = b_2(\gamma-\lambda)(q+a_0)$$

$$-[a_2(\gamma-\lambda)-\delta](q+b_0)$$

$$d_1 = \alpha[1+a_2(\gamma-\lambda)-\delta] - b_2(\gamma-\lambda).$$

This process may be stationary. A sufficient condition for stationarity is that $(1+\tau-\delta) < 1$, or that $\delta > \tau$. Obviously, the higher δ, the rate of depreciation of technology, the more likely it is that this condition will hold and output will be stationary. Thus, once depreciation is introduced, the value of the root of output may be greater than unity, less than unity, or not significantly different from unity, and even a model with endogenous technology may yield a stationary output process.

B. *A Real Business Cycle Model with Depreciation*

This subsection considers the effect of depreciation in a conventional business cycle model with exogenous technology. The model is that given by equation (22), but with one modification, namely, that the "permanent" technology shocks, the ξ's, dissipate slowly over time, so that equation (2) is replaced by

$$(30)\quad f_t = \bar{f}_t + \eta_t$$

$$\bar{f}_t = (1-\delta)\bar{f}_{t-1} + \xi_t,$$

and δ is again the depreciation rate. Using (30) to solve backward for $\bar{f}_{t-1}$ and substituting the result into the output equation (22) yields a stationary moving-average process for output:

$$(31)\quad y_t = q + b_0 + \eta_t + \xi_t$$

$$+ b_4 \sum_{j=0}^{\infty} (1-\delta)^j \xi_{t-j-1}.$$

Shocks to output have a declining impact over time. Because $(1-\delta) < 1$, the output process represented by this equation is stationary and has a less-than-unit root, unlike the model given by equation (22) without depreciation, which has a unit root.

While it is admittedly difficult to make a plausible case for the depreciation of technology in general, one cannot rule out the fact that some depreciation may occur in particular fields of knowledge. If technology does depreciate (δ is significantly greater than zero), conventional real business cycle models yield stationary output processes and cannot account for the observed nonstationarity of output or for persistence of innovations in output; only models with endogenous technology might display nonstationarity under these circumstances.

IV. Real Wages and Endogenous Technology

Empirical studies have generally failed to find significant countercyclical movements in the real wage. The most recent major study (Michael Keane et al., 1988) finds that real wages are mildly procyclical. This "stylized fact" is consistent with real business cycle models, but it appears to reject models with nominal wage contracts since real wages move countercyclically in such models.[14]

This conclusion—that wage-contracting models are rejected by studies of real wage movements—is less easily drawn if technology is endogenous. Even if the money wage is fixed, real wages may not exhibit pronounced countercyclical movement, particularly if learning occurs rapidly in some markets.

The general model presented above was not constructed to explain the cyclical pattern of real wages. Nonetheless, examination of the simple labor market model postulated in Section I can yield some insights. The only change necessary is to introduce overtime rates and to assume that at the beginning of the period overtime rates are also set in money terms. In Figure 1, at an expected price level P_0, a rise in employment above L_0 would result in overtime payments and raise the real and money wage. Hence the labor supply function is upward sloping, the slope depending on the structure of overtime pay.

A monetary shock that results in an unexpected rise in the price level reduces the real wage and the value of overtime payments, effectively shifting the labor supply schedule down and moving the labor market "equilibrium" from A to B. Employment (measured in man hours) rises: money wages rise through overtime pay, but by less than prices, so that the real wage falls below (W_o/P_o). In a real business cycle model, an exogenous rise in productivity would shift the demand for labor function to the right (to L_1^d) and increase both employment and real wages. Price and wage flexibility ensures that the labor supply function does not shift when prices change, and the new labor market equilibrium is at point C.

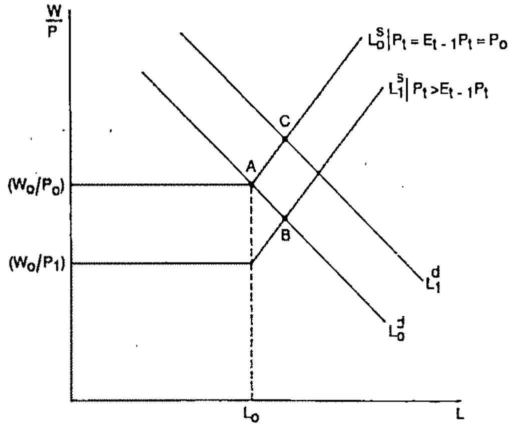


FIGURE 1. THE LABOR MARKET

The movement from A to B captures the conventional pattern of real wages and employment in a monetary business cycle model (in the Lucas-Barro models the rise in the labor supply results from misperception of the real return from working), but this pattern may not develop if technology depends on economic factors. In this case, as before, a monetary innovation boosts aggregate demand. Provided prices do not jump to a new higher level but adjust slowly (the excess money balances are spent gradually), then the labor supply function will begin to shift down slowly, raising employment. If the rise in employment results in a rapid increase in learning and innovation, the demand for labor will begin to rise. Thus, the downward movement of the labor supply schedule will set in motion an upward and rightward movement of the labor demand schedule. The resulting wage-employment pattern depends on the points of intersection of a downward-shifting labor supply function with an upward-shifting labor demand function. Rising prices reduce the real wage, but this is counteracted by rising productivity and output that place downward pressure on the price level and upward pressure on real wages; rising productivity retards the downward shift in the labor supply function. Clearly, a number of different wage paths are feasible, depending on the speed and magnitude of adjustment in the labor supply and demand functions in response to changing prices and productivity. It is conceivable that after a brief decline, real wages rise to their previous level. If there is a large increase in the demand for labor, one might even observe a *rise* in the real wage.

[14] This has led McCallum (1986) to suggest that price rather than wage stickiness must lie at the heart of cycles that are driven by aggregate demand shocks.

To conclude, real wage movements that are mildly procyclical but lagging *may* emerge from this model.[15] Once learning is introduced, the pattern of real wages over the business cycle becomes unclear: it need not be strongly countercyclical, even if nominal wages are fixed by contract.

V. Summary and Conclusions

Empirical work suggests that many innovations in technology take place in response to market conditions, particularly demand conditions. It was shown that, in a simple model of endogenous technical change, disturbances to output that originate on the supply side or the demand side of the economy have markedly different effects than if technology is assumed to be independent of the economic process.

Endogenous technical change acts as a propagation mechanism for shocks that have only transitory effects in conventional models. A supply-side shock, such as a temporary rise in productivity, or a demand-side shock, such as an unanticipated rise in aggregate demand, can induce a permanent upward shift in the aggregate production function. Changes on the supply side of the economy are not independent of changes on the demand side, and there is a long-run non-neutrality of money in models with endogenous technology. The models derived in this paper also imply that the growth rate of output will increase over time provided the supply of labor does not fall sufficiently to choke off rising growth rates.

It was shown that if technology depreciates, conventional real business cycle models cannot account for the observed nonstationarity of output. However, both real and monetary models with endogenous technology still yield nonstationary output processes provided the rate of depreciation of technical knowledge is not too high. Furthermore, the pattern of real wage movements is unclear if technical knowledge changes quite rapidly in response to changes in factor utilization. Even if money wages are fixed, real wages may display very little countercyclical movement.

There has been some debate as to whether models that emphasize aggregate demand innovations, or models that rely on real shocks as an impulse mechanism, provide a better characterization of output fluctuations. However, if technology is endogenous, real and monetary business cycle models exhibit very similar properties, as a comparison of equations (19) and (25) illustrates. Both classes of models can account for the persistence of innovations in output, and both can account for the rise in the growth rates of real output observed by Romer. This similarity makes it difficult, at the present time, to reject one model in favor of the other in terms of providing a better account of the output process.

REFERENCES

Arrow, Kenneth J., "The Economic Implications of Learning by Doing," *Review of Economic Studies*, June 1962, *29*, 155–73.

Barro, Robert J., "Rational Expectations and the Role of Monetary Policy," *Journal of Monetary Economics*, January 1976, *2*, 1–32.

______, "Rational Expectations and Macroeconomics in 1984," *American Economic Review, Papers and Proceedings*, May 1984, *24*, 179–82.

Bils, Mark J., "Real Wages over the Business Cycle: Evidence from the Panel Data," *Journal of Political Economy*, August 1985, *93*, 666–89.

Campbell, John Y. and Mankiw, N. Gregory, "Are Output Fluctuations Transitory?" *Quarterly Journal of Economics*, November 1987, *102*, 857–80.

Cochrane, John H., "How Big Is the Random Walk in GNP?" *Journal of Political Economy*, October 1988, *96*, 893–920.

Eltis, W. A., "The Determination of the Rate

[15] The argument is made more formally in Stadler (1989). A question that is not satisfactorily addressed in the literature is the degree to which real wage changes are lagging over the business cycle. Victor Zarnowitz comments that "Money wages often rise less than prices in recoveries and more than prices in late expansion stages. There is a procyclical and lagging movement in labor costs per unit output." [1985, p. 529]

of Technical Progress," *Economic Journal*, September 1971, *81*, 502–24.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–205.

Hamermesh, Daniel H. and Rees, Albert, *The Economics of Work and Pay*, New York: Harper & Row, 1984.

Kamien, Morton and Schwartz, Nancy J., *Market Structure and Innovation*, Cambridge: Cambridge University Press, 1982.

Keane, Michael, Moffit, Robert and Runkle, David, "Real Wages over the Business Cycle: Estimating the Impact of Heterogeneity with Micro Data," *Journal of Political Economy*, December 1988, *96*, 1232–66.

Kennedy, Charles and Thirlwall, A. P., "Technical Progress: A Survey," *Economic Journal*, March 1972, *82*, 11–72.

King, Robert G., Plosser, Charles I. and Rebelo, Sergio T., (1988a) "Production, Growth and Business Cycles: I. The Basic Neoclassical Model," *Journal of Monetary Economics*, March 1988, *21*, 195–232.

______, ______, and ______, (1988b) "Production, Growth and Business Cycles: II. New Directions," *Journal of Monetary Economics*, May 1988, *21*, 309–41.

______ and Rebelo, Sergio T., "Business Cycles with Endogenous Growth," unpublished paper, 1986.

Kydland, Finn and Prescott, Edward C., "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Long, John B. and Plosser, Charles I., "Real Business Cycles," *Journal of Political Economy*, February 1983, *91*, 39–69.

Lucas, Robert E., Jr., "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

______, "On the Mechanics of Economic Development," *Journal of Monetary Economics*, July 1988, *22*, 3–42.

McCallum, Bennett T., "On 'Real' and 'Sticky-Price' Theories of the Business Cycle," *Journal of Money, Credit, and Banking*, November 1986, *18*, 397–414.

Nelson, Charles R. and Plosser, Charles I., "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, September 1982, *10*, 139–62.

Prescott, Edward C. and Boyd, John H., "Dynamic Coalitions: Engines of Growth," *American Economic Review, Papers and Proceedings*, May 1987, *77*, 63–67.

Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*, 1002–37.

Schmookler, Jacob, *Invention and Economic Growth*, Cambridge, MA: Harvard University Press, 1966.

Stadler, George W., "Real Versus Monetary Business Cycle Theory and the Statistical Characteristics of Output Fluctuations," *Economics Letters*, 1986, *22*, 51–54.

______, "Models of Business Cycles with Endogenous Technology," unpublished doctoral thesis, University of Western Ontario, 1988.

______, "Real Wages over the Business Cycle with Nominal Wage Contracting," unpublished paper, 1989.

Taylor, John B., "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, February 1980, *88*, 1–23.

Utterback, James, M., "Innovation in Industry and the Diffusion of Technology," *Science*, February 1974, *183*, 620–26.

Watson, Mark W., "Univariate Detrending Methods with Stochastic Trends," *Journal of Monetary Economics*, July 1986, *18*, 1–27

West, Kenneth D., "On the Interpretation of Near Random-Walk Behavior in GNP," *American Economic Review*, March 1988, *78*, 202–209.

Zarnowitz, Victor., "Recent Work on Business Cycles in Historical Perspective: A Review of Theories and Evidence," *Journal of Economic Literature*, June 1985, *23*, 523–80.

The Economist, March 21, 1987, *302*.

______, April 16, 1988, *307*.

A Theory of Managed Trade

By Kyle Bagwell and Robert W. Staiger*

This paper proposes a theory that predicts low levels of protection during periods of "normal" trade volume coupled with episodes of "special" protection when trade volumes surge. This dynamic pattern of protection emerges from a model in which countries choose levels of protection in a repeated game facing volatile trade swings. High trade volume leads to a greater incentive to defect unilaterally from cooperative tariff levels. Therefore, as the volume of trade expands, the level of protection must rise in a cooperative equilibrium to mitigate the rising trade volume and hold the incentive to defect in check. (JEL *411, 422*)

Two major trends have dominated the postwar history of trade policy in industrailized countries. One is the dramatic multilateral reduction in tariffs negotiated under the General Agreement on Tariffs and Trade (GATT).[1] The other is the move toward "special" protection that has occurred as the industrialized countries of the world have become more integrated and as volatility in trade flows has become a more important source of domestic disruption. The rise in special forms of protection is epitomized by the growing use of Voluntary Export Restraints (VERs), Orderly Market Arrangements (OMAs), and tariffs that are tailor-made to suit the needs of particular sectors.[2] These policy tools are typically utilized by countries to limit the rate of expansion of imports or exports from that which would occur absent intervention. The term "managed trade" is often invoked to characterize the current international trading environment, since it consists of a relatively low "baseline" or "normal" level of protection combined with the use of special protection to dampen underlying changes in trade flows.

The low baseline level of protection sustained by countries suggests that a static noncooperative Nash equilibrium view is inadequate to explain existing levels of protection. One alternative is to view the existing trading environment as the result of explicit agreements among countries. This approach to explaining levels of protection has been taken by Wolfgang Mayer (1981) and Raymond Riezman (1982). However, such explicit agreements require the existence of a workable enforcement mechanism, and at the international level it is unclear what that mechanism might be. A second alternative is to consider only self-enforcing agreements or tacit cooperation among countries. As Avinash Dixit (1987) and Richard Jensen and Marie Thursby (1984) have shown, the (credible) threat of future punishment can sustain a more liberal trading environment than that predicted under the static Nash equilibrium. These models can help explain how countries are able to sustain a relatively liberal trading environment in "normal" periods. However, they remain silent on the issue of "special" protection, since they take each

*Kyle Bagwell, Department of Economics, Northwestern University, Evanston, IL 60201; Robert W. Staiger, Department of Economics, Stanford University, Stanford, CA 94305; and National Bureau of Economic Research. We thank K. C. Fung, Marie Thursby, and participants in the International Seminars at Berkeley, Stanford, UCLA, the Institute for International Economic Studies in Stockholm, and the NBER Summer Institute for very helpful comments. We are especially indebted to two anonymous referees for extremely helpful suggestions. All remaining errors are ours.

[1] In the United States, for example, the average tariff level on dutiable imports (the ratio of duties collected to dutiable imports) fell from 53.5 percent in 1933 to 5.2 percent in 1982 (U.S. Trade Representative, 1984, p. 187).

[2] The increasing relative importance of special protection as an instrument of international trade policy has been widely noted. See, for example, S. A. B. Page (1979) and C. Fred Bergsten and William Cline (1983) for attempts to quantify this trend.

period to be the same as every other. In this regard, W. Max Corden (1974, pp. 175–76) has argued that countries rarely initiate protection for the purpose of capturing terms-of-trade *gains*, presumably because of the fear of future retaliation by their trading partners, but that countries do employ protection for its terms-of-trade effects in periods when their terms-of-trade would otherwise *decline*. Corden argues that retaliation by trading partners is less likely during such periods. This suggests that episodes of "special" protection might usefully be viewed as part of a tacit international agreement in a *changing* environment.

We attempt to formalize this view by considering the way in which sustainable levels of protection in tacit cooperative equilibria are affected by changes in the underlying trade volume. Since potentially exploitable terms-of-trade effects will embody greater potential national welfare gains the greater is the underlying volume of trade, periods of high trade volume are likely to correspond to periods of great incentive to exercise one's power over the terms-of-trade. If the trade volume is large enough, the immediate gains from protection may outweigh the losses from punishment, and free trade will be unsustainable. However, this does not imply that international cooperation need break down. Countries can cooperatively utilize protection during periods of exceptionally high trade volume to mitigate the incentive of any country to unilaterally defect, and in so doing can avoid reversion to the noncooperative Nash equilibrium. Thus, surges in the underlying trade volume lead to periods of "special" protection as countries attempt to maintain some level of international cooperation. In this sense, the model we develop below depicts managed trade as the outcome of tacit cooperation among countries in the presence of volatile trade swings.

We adopt a very simple partial equilibrium framework within which to make these points. Section I lays out the basic model under the assumption of free trade and calculates the underlying free trade volume as a function of the parameters of the model. Section II solves for the static Nash equilibria in the quota and tariff games. Our results here are similar to those developed in Dixit (1987). These equilibria constitute the credible (subgame perfect) punishments in the dynamic game of the following section, the threat of which will be used to support tacit cooperation. The dynamic model for the quota and tariff games is analyzed in Section III, where it is shown that equilibrium trade policy becomes more restrictive during periods of high (free) trade volume. This result is reminiscent of a related point made by Julio Rotemberg and Garth Saloner (1986a) in the context of a repeated oligopoly model with demand shocks.[3] Our modeling approach is clearly inspired by their work. Section IV adds a second sector and considers the model's implications for the relationship between bilateral trade imbalances and protection. Section V discusses the generality of our results and considers several extensions. Section VI offers some conclusions.

I. Free Trade

We begin with the characterization of free trade in a very simple partial equilibrium model of trade in a single sector between two countries. Let the world (two-country) output in the sector be fixed at 2. At the beginning of any period, the distribution of world output between the "domestic" (no *) and the "foreign" (*) country is determined by a commonly known distribution function $F(e)$ that generates foreign output $e \in [0,1]$, with domestic output then given by $2-e$. On the demand side, the domestic and foreign countries are assumed to have identical linear demands $C=\alpha-\beta P$ and $C^*=\alpha-$

[3] In concluding, Rotemberg and Saloner (1986a) conjecture that their framework could yield a prediction of trade wars occurring in states of depressed demand. Our paper addresses a different issue, but in a similar spirit. Also, Riezman (1987) introduces random terms-of-trade shocks into the tariff model of Dixit (1987) and notes that shocks to the terms-of-trade that increase the current gain from defection will increase the likelihood of Nash reversion. His concern, however, is with the effect of unobservability of shocks and tariffs on the ability of countries to sustain low cooperative tariffs.

βP^*, respectively, where $C(C^*)$ is the consumption level of the domestic (foreign) country and $P(P^*)$ is the domestic (foreign) country price. Competitive firms supply the product in each country. For simplicity, we assume that production costs are zero and that $\alpha > 2$.

Free trade will ensure that a single price P^f prevails in both markets so that $P = P^* = P^f$. The equilibrium condition that world supply equals world demand, $2 = C(P^f) + C^*(P^f)$, determines the free trade price P^f as $P^f = (\alpha - 1)/\beta$. Thus, consumption levels under free trade are given by $C(P^f) = C^*(P^f) = 1$. Finally, the free trade volume V^f is given by $V^f = 2 - e - C(P^f) = C^*(P^f) - e = 1 - e$. In periods when $e = 1$, both countries have equal supply and there will be no trade between them. When $e < 1$, the domestic country exports the quantity $1 - e$. Hence, the domestic (foreign) country is the exporter (importer), and free trade volume rises as e falls away from 1. This completes the characterization of trade volume under conditions of free trade.

II. A Static Model of Protection

In this section we characterize the set of static Nash equilibria for the simple model of the previous section when countries choose either trade taxes or quotas. These equilibria will serve as credible (subgame perfect) punishments in the dynamic games considered in the next section, the threat of which can support tacit cooperation in a repeated setting.

The two countries are assumed to observe the current realization of e, and thus the trade volume that would prevail in the period under free trade, $V^f = 1 - e$, and then to choose simultaneously their protective policies for the period.[4] The domestic country is thus choosing export taxes or quotas while the foreign country chooses corresponding import policies.[5] Each country's objective is to maximize its sum of producer surplus, consumer surplus, and rents from protection.[6]

Consider first the determination of the static Nash equilibrium when countries choose trade quantities directly in the form of import and export quotas. Provided that quota licenses are either auctioned off by the governments of each country or simply given to their respective firms, the country whose quota binds—the country with the smaller quota—will capture all the quota rents. This means that as long as there is trade, each country can always do better by tightening its quota beyond that set by its trading partner, which leads to the well-known property that the unique static Nash equilibrium in the quota game is autarky (see, for example, Edward Tower, 1975). Thus, regardless of the realization of e and the underlying free trade volume V^f, the static Nash equilibrium in the quota game ensures that no trade will take place.

We turn now to the determination of the static Nash equilibrium level of protection when the domestic and foreign countries choose specific export and import taxes $\tau(V^f)$ and $\tau^*(V^f)$, respectively, as functions of the observed free trade volume.[7] To begin, the actual volume of trade following the realization of the free trade volume V^f and the selection of trade taxes $\tau(V^f)$ and $\tau^*(V^f)$ is easily characterized. If trade occurs, then effective prices to producers in the domestic (exporting) country must be equal across countries and world supply and demand must also be equal. Since (nonnegative) taxes cannot reverse the free di-

[4] Our results would not be significantly altered if countries had common and imperfect information about V^f when choosing policies. If, however, each country had some private information about the current V^f, then the analysis would be considerably more complex. For an analysis of the role of private information in finitely repeated tariff games, see Jensen and Thrusby (1989).

[5] We discuss in Section V an extension of our results to two sectors in which only import-restricting policies can be used.

[6] Countries are not concerned with risk sharing in this partial equilibrium setting, since national income, and thus the marginal utility of national income, is unaffected by shocks in this single sector. See Section V for a discussion of the case where sectoral income smoothing motivates trade policy intervention.

[7] Our results are essentially unchanged if countries set *ad valorem* tariffs.

rection of trade, we have $P^* - P = \tau + \tau^*$ and $2 = C(P) + C^*(P^*)$. Trade will occur when trade taxes are not prohibitive; in our model, it is easy to show that trade occurs provided

$$(1) \qquad 2V^f/\beta > \tau + \tau^*.$$

Now, assuming that (1) holds, we have $P(\tau,\tau^*) = (\alpha-1)/\beta - (\tau+\tau^*)/2$ and $P^*(\tau,\tau^*) = (\alpha-1)/\beta + (\tau+\tau^*)/2$, which gives prices as functions of trade taxes for the domestic and foreign country.[8]

Letting $W(V^f,\tau,\tau^*)$ and $W^*(V^f,\tau,\tau^*)$ represent domestic and foreign country welfare, respectively, given by the sum of the country's consumer surplus, producer surplus, and tariff revenue, we have when (1) holds

$$(2) \quad W(V^f,\tau,\tau^*) = \int_{P(\tau,\tau^*)}^{\alpha/\beta} C(P)\,dP + \int_0^{P(\tau,\tau^*)} [1+V^f]\,dP + \tau X(V^f, P(\tau,\tau^*)),$$

$$(3) \quad W^*(V^f,\tau,\tau^*) = \int_{P^*(\tau,\tau^*)}^{\alpha/\beta} C^*(P^*)\,dP^* + \int_0^{P^*(\tau,\tau^*)} [1-V^f]\,dP^* + \tau^* M(V^f, P^*(\tau,\tau^*)),$$

where we have written each country's output in terms of the underlying free trade volume V^f, and where $X(V^f, P(\tau,\tau^*))$ and $M(V^f, P^*(\tau,\tau^*))$ are domestic export supply and foreign import demand, respectively, written as functions of V^f and tariff-distorted prices.

The remaining possibility is that (1) fails. Now trade taxes prohibit trade. This possibility corresponds to autarky, with domestic and foreign prices given, respectively, by $P(V^f) = (\alpha - 1 - V^f)/\beta$ and $P^*(V^f) = (\alpha - 1 + V^f)/\beta$ and welfare

$$(4) \quad W(V^f,\tau,\tau^*) = \int_{P(V^f)}^{\alpha/\beta} C(P)\,dP + \int_0^{P(V^f)} [1+V^f]\,dP,$$

$$(5) \quad W^*(V^f,\tau,\tau^*) = \int_{P^*(V^f)}^{\alpha/\beta} C^*(P^*)\,dP^* + \int_0^{P^*(V^f)} [1-V^f]\,dP^*.$$

With the payoff functions now defined by (2), (3), (4), and (5), a *Nash equilibrium* for the static tariff game can be defined as a pair of tariff functions, $\tau_N(V^f)$ and $\tau_N^*(V^f)$, such that for every $V^f \epsilon [0,1]$, $\tau_N(V^f)$ maximizes $W(V^f,\tau,\tau_N^*(V^f))$ over τ and $\tau_N^*(V^f)$ maximizes $W^*(V^f,\tau_N(V^f),\tau^*)$ over τ^*.

To solve for Nash equilibria, we first characterize best-response correspondences, $\tau_R(V^f,\tau^*)$ and $\tau_R^*(V^f,\tau)$, defined, respectively, as the maximizers of $W(V^f,\tau,\tau^*)$ and $W^*(V^f,\tau,\tau^*)$. If (1) holds, $dW(V^f,\tau,\tau^*)/d\tau = V^f/2 - (3\beta\tau + \beta\tau^*)/4$, and $W(V^f,\tau,\tau^*)$ is strictly concave in τ. Hence,

$$(6) \quad \tau_R(V^f,\tau^*) = 2V^f/3\beta - \tau^*/3, \qquad \text{if } \tau^* < 2V^f/\beta.$$

If instead $\tau^* \geq 2V^f/\beta$, then by (1) any τ generates autarky. The domestic country welfare is then independent of its tariff, and so

$$(7) \quad \tau_R(V^f,\tau^*) = [0,\infty), \text{ if } \tau^* \geq 2V^f/\beta.$$

Similar calculations for the foreign country

[8] Note that as long as tariffs are not prohibitive so that (1) holds, prices will only be affected indirectly by the realization of e, through τ and τ^*. The absence of a direct impact of e on prices is due to the perfect negative correlation across countries of the supply shocks we consider. While a useful simplifying assumption, we argue in Section V that our results will be preserved in much more general models.

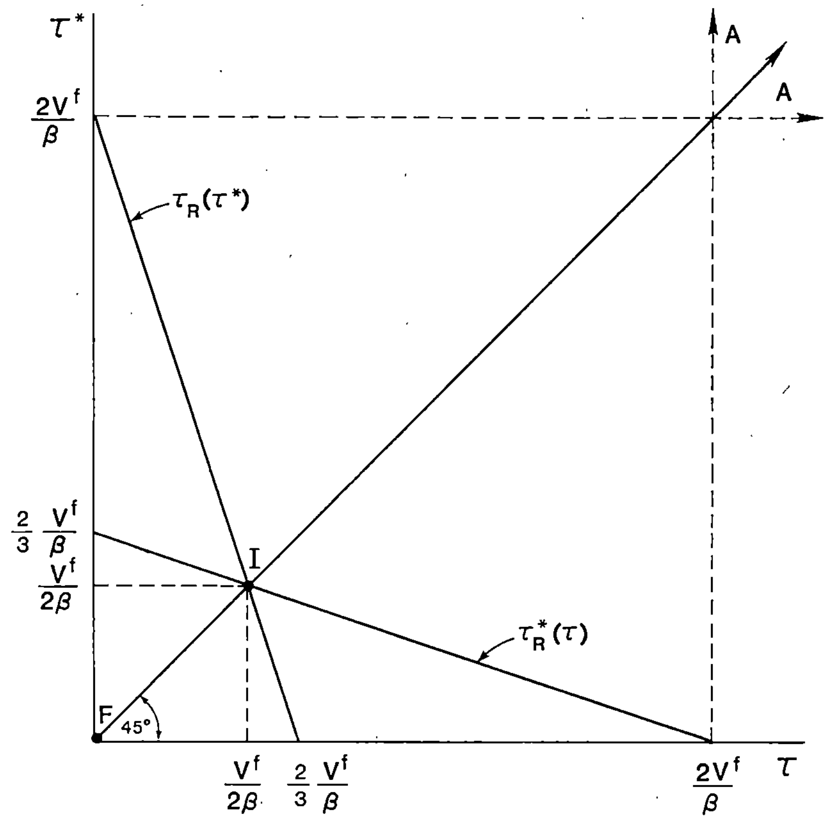


FIGURE 1

give the exactly symmetric best-response correspondence $\tau_R^*(V^f,\tau)$.

There are thus two disjoint sets of Nash equilibria. The interior equilibrium, found by solving (6) and its symmetric counterpart, has $\tau_N(V^f) = V^f/2\beta = \tau_N^*(V^f)$. The other equilibrium set corresponds to autarky. Any (τ, τ^*) such that $\tau \geq 2V^f/\beta$ and $\tau^* \geq 2V^f/\beta$ forms a no-trade Nash equilibrium with $\tau_N(V^f) = \tau$ and $\tau_N^*(V^f) = \tau^*$ for all V^f. Figure 1 illustrates the sets of Nash equilibria for the static tariff game.

To summarize, the static tariff game between countries generates two sets of equilibria: an interior Nash equilibrium and a set of autarky Nash equilibria. The static quota game has autarky as its unique Nash equilibrium.

Finally, we note that the equilibrium payoff configurations in the static tariff and quota games can be unambiguously ranked. In all (tariff and quota) autarky equilibria, payoffs are given by (4) and (5). Setting τ and τ^* in (2) and (3) first equal to zero and then equal to the Nash interior tariffs, it is straightforward to verify that welfare is highest for both countries under free trade and higher in the interior Nash equilibrium than in the autarky Nash equilibria if and only if $V^f > 0$.[9]

[9] Harry Johnson (1953–54) notes that one country may prefer the (internal) Nash tariff equilirium to free trade if its import demand is sufficiently elastic relative to that of its trade partner. John Keenan and Riezman (1988) have linked this possibility to differences in country size. While the symmetry in our model, which results in common interior Nash tariffs, does not allow this possibility to arise, the main complication it would introduce to our analysis is to alter the focus from symmetric to asymmetric tariff equilibria. See also the discussion in Section V.

III. A Dynamic Model of Protection

We now extend the model to allow for repeated interaction. In particular, we explore the sense in which a dynamic environment enables countries to lower protection from the levels that would prevail in a static setting and characterize the relation between the achieved protection and trade volume.

The dynamic game upon which we focus is simply the static game studied above infinitely repeated. Thus, at the start of any period, a value for e (and thereby V^f) is realized and observed by all. Current period protection policies are then set, and current welfare is determined. At the beginning of the next period, all past choices are observed and a new value for e (and thus V^f) is determined. We assume that e is drawn from the same distribution independently every period.

We examine symmetric (subgame perfect) Nash equilibria, in which the countries cooperate with low, common protection levels and credibly threaten to forever revert to a static Nash equilibrium if cooperation is violated. In the tariff game, common cooperative tariff levels imply that both countries share symmetrically in the cooperative tariff rents. Analogously, in the quota game, we assume that quota rents are shared symmetrically in the cooperative equilibrium.[10]

As discussed above, the most preferred symmetric trade policy is free trade, and lower symmetric levels of protection are always preferred jointly to higher symmetric levels of protection. For some values of V^f, we will see that the threat of reversion is sufficient to generate free trade. However, for other values of V^f, free trade cannot be maintained and the cooperative level of protection entails positive symmetric tariffs.

We consider first the tariff game. The cooperative trade tax function, $\tau_c = \tau_c(V^f)$, must provide each country with no incentive to defect. That is, for every V^f, the expected discounted welfare to each country under the strategy $\tau_c(V^f)$ must be no less than the welfare achieved by the country when defecting and thereafter receiving the expected discounted welfare associated with a static Nash equilibrium. Clearly, a country choosing to defect does best by selecting a tariff on its reaction curve. Thus, from (6), if countries are cooperating and allowing trade, the optimal tariff with which to defect is

$$(8)\qquad \tau_D(V^f,\tau_c) = 2V^f/3\beta - \tau_c/3 = \tau^*_D(V^f,\tau_c).$$

We now fix V^f and a cooperative tariff level τ_c and characterize the static incentive to defect. Let

$$(9)\qquad \Omega(V^f,\tau_D(V^f,\tau_c),\tau_c) \equiv W(V^f,\tau_D(V^f,\tau_c),\tau_c) - W(V^f,\tau_c,\tau_c)$$

$$(10)\qquad \Omega^*(V^f,\tau_c,\tau_D(V^f,\tau_c)) \equiv W^*(V^f,\tau_c,\tau_D(V^f,\tau_c)) - W^*(V^f,\tau_c,\tau_c)$$

represent the respective static gains from defection for the domestic and the foreign country.

Figure 2 illustrates the static incentive to defect from the cooperative tariff $\tau_c \geq 0$. The foreign import demand curve is given by the downward sloping line with slope $-1/\beta$. The domestic export supply curve is given by the upward sloping line with slope $1/\beta$. The cooperative tariff τ_c results in prices $P^*(\tau_c,\tau_c)$ and $P(\tau_c,\tau_c)$ in the foreign and the domestic country, respectively. The

[10] It is natural to focus on common protection levels in this simple model, where countries are assumed symmetric. Moreover, one can show that symmetric rent sharing supports the highest degree of cooperation in this model. A more general model is discussed in Section V, where our basic conclusions hold but asymmetries play a real role. See also Robert Feenstra and Tracy Lewis (1987), who find in a different context that sharing the rents from protection allows countries to avoid trade wars.

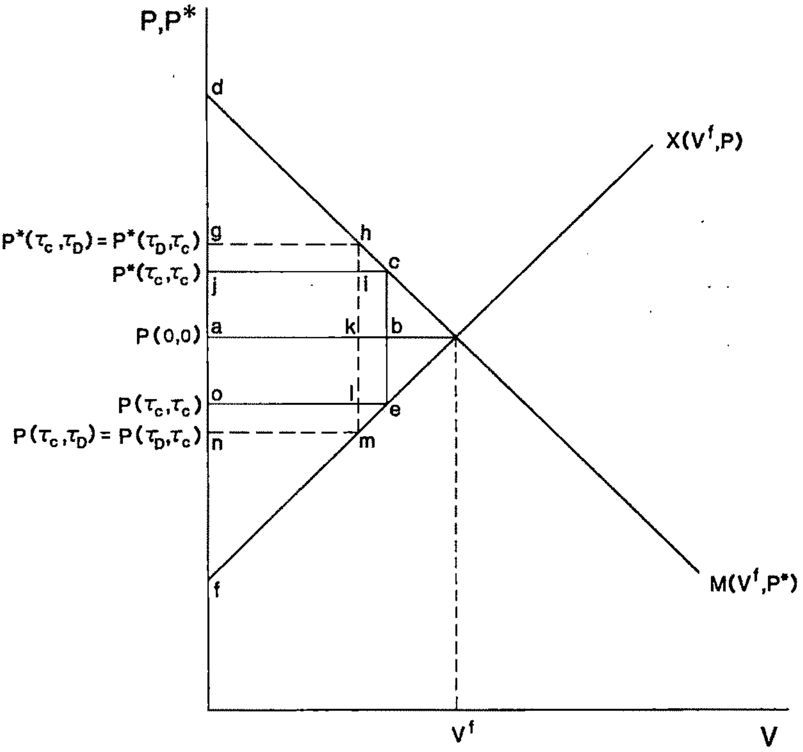


FIGURE 2

foreign gain from trade under τ_c is the sum of the surplus from trade, given by the area of the triangle *djc*, and the foreign share of cooperative tariff revenues, given by the area of the rectangle *jcba*. The domestic gain from trade under τ_c is analogously given by the sum of the areas of the triangle *oef* and the rectangle *oeba*.

Now consider a defection from τ_c. We know from (8) that both countries will defect to the same τ_D. Hence, whichever country defects, prices will be $P^*(\tau_c,\tau_D) = P^*(\tau_D,\tau_c)$ and $P(\tau_c,\tau_D) = P(\tau_D,\tau_c)$ in the foreign and domestic countries, respectively. The foreign static gain from defection is then given by the net increase in its collection of tariff revenues, represented by the difference between the rectangles *onml* and *kbci*, minus the efficiency loss in its surplus from trade, represented by the triangle *hic*. Analogously, the domestic exporter's static gain from defection is given by *gjih* – *kbel* – *mle*. As is evident from Figure 2, the equivalence

$$\Omega\big(V^f,\tau_D(V^f,\tau_c),\tau_c\big)=\Omega^*\big(V^f,\tau_c,\tau_D(V^f,\tau_c)\big)$$

holds for all $\tau_c \geq 0$. All results concerning the static incentive to defect can therefore be expressed in the domestic country notation.

Using the envelope theorem, we find that

$$\text{(11)}\quad \frac{d\Omega\big(V^f,\tau_D(V^f,\tau_c),\tau_c\big)}{dV^f} = \big[\tau_D(V^f,\tau_c)-\tau_c\big]/2,$$

$$\text{(12)}\quad \frac{d\Omega\big(v^f,\tau_D(V^f,\tau_c),\tau_c\big)}{d\tau_c} = \frac{\beta}{4}\big[5\tau_c-\tau_D(V^f,\tau_c)\big]-\frac{V^f}{2}.$$

Using (8), we see that $\Omega(V^f, \tau_D(V^f, \tau_c), \tau_c)$ is strictly increasing in V^f and strictly decreasing in τ_c if and only if

$$(13) \qquad \tau_c < V^f/2\beta.$$

Provided that the cooperative tariff is below the static interior Nash tariff, the incentive to defect from a fixed τ_c is larger the larger is V^f and the smaller is τ_c.

These conditions are simple to interpret. As the underlying free trade volume increases, the incentive to defect gets larger. This occurs because the terms-of-trade gains from defection are applied to a larger trade volume, that is, because more tariff revenue is collected from one's trading partner under defection when the underlying trade volume is high. The incentive to defect can be mitigated by increasing the cooperative tariff, which acts to reduce the volume of trade. Thus, when the trade volume surges, one might suspect that a high τ_c would be required to avoid defection. This is indeed what we will find.

Having characterized the static incentive to defect, our next step is to examine the expected future loss suffered by a country that defects. Letting E be the expectations operator with expectations taken over V^f and δ be the discount factor, we represent the respective present discounted values of the expected future gain from not defecting today for the domestic and the foreign country as

$$(14) \quad \frac{\delta}{1-\delta}\big[EW(V^f, \tau_c(V^f), \tau_c(V^f)) - EW(V^f, \tau_N(V^f), \tau_N(V^f))\big] \equiv \omega(\tau_c(\cdot)),$$

$$(15) \quad \frac{\delta}{1-\delta}\big[EW^*(V^f, \tau_c(V^f), \tau_c(V^f)) - EW^*(V^f, \tau_N(V^f), \tau_N(V^f))\big] \equiv \omega^*(\tau_c(\cdot)).$$

Since e (and thus V^f) is i.i.d. across periods, ω and ω^* are independent of the current value of V^f as well as the current value of $\tau_c(V^f)$. The function $\tau_c(\cdot)$ will affect ω and ω^*, however, since the function's distributional characteristics influence the corresponding expected values. Observe that ω and ω^* will be strictly positive when $\delta > 0$ and $\tau_c(V^f) < \tau_N(V^f)$ for all V^f, in which case the threat of future punishment is meaningful.

We focus on the case where punishment involves infinite reversion to the interior Nash tariff equilibrium of the static game.[11] Using (2) and (3), we then have

$$(16) \quad \omega(\tau_c(\cdot)) = \omega^*(\tau_c(\cdot)) = \frac{\delta}{(1-\delta)}\left[\frac{\sigma_v^{2f} + (EV^f)^2}{8\beta} - \frac{\beta}{2}\left[\sigma_{\tau_c}^2 + (E\tau_c(V^f))^2\right]\right],$$

where σ_v^{2f} is the variance of V^f and $\sigma_{\tau_c}^2$ is the variance of $\tau_c(V^f)$. Note that the expected future gain from cooperating is higher when σ_v^{2f} and EV^f are higher, holding fixed $\tau_c(V^f)$.[12] This reflects the fact that the gains associated with cooperation (low protection) are increasing and convex in the underlying free trade volume.

[11] The case of reversion to autarky is similar, as detailed in Bagwell and Staiger (1988). The autarky threat actually generates the most cooperation possible, since it is the harshest punishment that countries will endure. (This point is made at a general level by Dilip Abreu, 1988). Yet, its very severity makes it somewhat less plausible than the threat to revert to the interior equilibrium, As Dixit (1987) notes, the possibility of autarky reversion is also significant for the tariff game since the existence of two static Nash equilibria makes possible cooperation even in finite horizon settings. (See Jean-Pierre Benoit and Vijay Krisna, 1985, and James Friedman, 1985, for more general discussions of this issue.)

[12] The assumption of linear demand generates a welfare function that is quadratic in τ. Since the Nash tariff is linear in V^f, only the mean and variance of V^f appear in (16). The higher-order moments might be important with nonlinear demand.

We have now characterized both the immediate gain from defection and the expected future loss. Both expressions are identical across countries, which enables us to focus on the domestic country henceforth. Now for credible cooperation to occur, the cooperative trade tax function, $\tau_c(V^f)$, must be such that at every V^f, no country has incentive to defect, or

$$(17)\quad \Omega\big(V^f,\tau_D\big(V^f,\tau_c(V^f)\big),\tau_c(V^f)\big) \le \omega\big(\tau_c(\cdot)\big).$$

This is our fundamental "no defection" condition, which implicitly defines a cooperative trade tax function.

There will in general be many functions that satisfy (17). To characterize the "most cooperative" trade tax function, we hold ω fixed at a constant level and solve for the lowest, nonnegative τ_c satisfying (17).[13] This process generates a trade tax function, with independent variables V^f and ω, and determines the necessary properties that the most cooperative trade tax function must satisfy.

To begin, fix $\omega > 0$. For $V^f = 0$, (8) and (9) establish that $\Omega(0,\tau_D(0,0),0) = 0$, so that (17) is satisfied by $\tau_c = 0$. Holding τ_c fixed at zero and increasing V^f, we know from (11) that $\Omega(V^f,\tau_D(V^f,0),0)$ increases monotonically. If ω is not too large, which will always be the case if δ is not too large, then there exists a critical value of $V^f, \bar{V}^f$, such that $\Omega(\bar{V}^f,\tau_D(\bar{V}^f,0),0) = \omega$.

Solving this explicitly gives

$$(18)\qquad \bar{V}^f = \sqrt{6\beta\omega}\,.$$

Hence, free trade is sustainable for $V^f \in [0,\bar{V}^f]$.

For more extreme values of V^f, where $V^f \in (\bar{V}^f,1]$, (17) will be violated at $\tau_c = 0$. From (12), τ_c must then rise above zero to reestablish (17). Explicit calculation yields the following representation of the most cooperative tax rule:

$$(19)\quad \tau_c(V^f,\omega) = \begin{cases} 0, \text{ if } V^f \in [0,\bar{V}^f] \\ \dfrac{V^f-\bar{V}^f}{2\beta} \text{ if } V^f \in [\bar{V}^f,1] \end{cases}.$$

The corresponding cooperative trade volume is given by

$$(20)\quad V^c = \begin{cases} V^f \text{ if } V^f \in [0,\bar{V}^f] \\ \dfrac{V^f+\bar{V}^f}{2} \text{ if } V^f \in [\bar{V}^f,1] \end{cases}.$$

The next figure summarizes (19) and (20). Figure 3 plots τ_c and V^c as a function of V^f. The threat to revert to the static interior Nash equilibrium supports free trade over a range of moderate trade volume. Intuitively, if the underlying free trade volume is low, then the (current) incentive to defect with a tariff is low, even if the natural flow of trade is unrestricted ($\tau_c = 0$). Once the trade volume becomes extreme, the trade tax function increases with the magnitude of the volume. Here, the incentive to defect is large because the natural flow of trade is high, and so the volume of trade must be mitigated somewhat ($\tau_c > 0$, $V^c < V^f$) in order to prevent defection.

The above analysis characterizes the necessary features of the optimal trade tax function and gives us an expression $\tau = \tau_c(V^f,\omega)$. The analysis was conducted under the assumption of an exogenously given ω. In fact, as (16) illustrates, ω depends on the $\tau_c(\cdot)$ function, $\omega = \omega(\tau_c(\cdot))$. To establish existence of the optimal trade tax function, we must ensure that these equations are consistent, so that the ω with which we began is also the ω value that $\tau_c(V^f,\omega)$ generates. Substituting the first equation into the second and using (16), (18), and (19), we can write the resulting equations as $\tilde{\omega}(\omega) = \omega$, since $\omega(\tau_c(\cdot))$ is independent of V^f. The

[13] A focus on the most cooperative equilibrium seems natural here, since countries are free to communicate about which self-enforcing equilibrium they will settle on, and the most cooperative equilibrium is the only symmetric equilibrium that is not Pareto dominated. Indeed, the GATT may be viewed as a forum within which the most cooperative (self-enforcing) trading arrangements are codified.

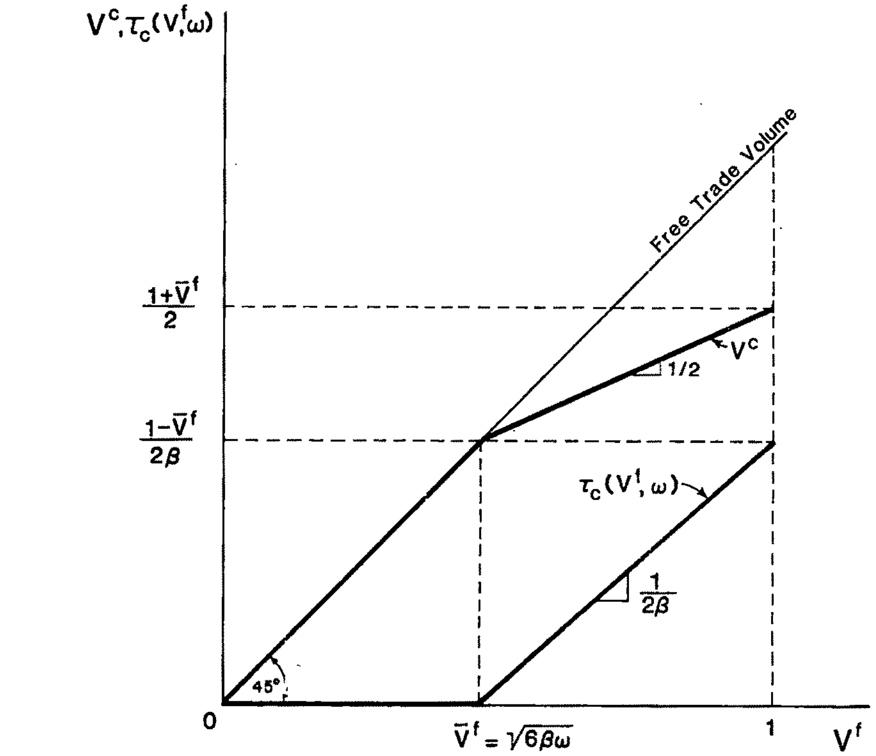


FIGURE 3

most cooperative trade tax function can then be represented as $\tau_c = \tau_c(V^f)$, when the largest $\hat{\omega}$ such that $\hat{\omega} \in (0, 1/6\beta)$ and $\tilde{\omega}(\hat{\omega}) = \hat{\omega}$ is substituted into $\tau_c(V^f, \omega)$.

We must now prove that such a fixed point does exist. Observe first that a fixed point does occur at $\hat{\omega} = 0$, corresponding to the continual play of the static interior equilibrium. This follows since $\tau_c(V^f, 0) = \tau_N(V^f)$, by (19). To explore the possibility of a positive root, we explicitly calculate $E\tau_c(V^f)$ from (18) and (19) and use (16) to get

$$(21)\quad \tilde{\omega}(\omega) = \frac{\delta}{8\beta(1-\delta)}\left[\sigma_v^{2f} + (EV^f)^2 - \int_{\bar{V}^f}^{1} (V^f - \bar{V}^f)^2 dF(V^f))\right],$$

if $\omega \in [0, 1/6\beta]$, where $F(V^f)$ is the distribution function for V^f. It is now straightforward to verify that with primes denoting derivatives, $\tilde{\omega}(\omega = 0) = 0$, $\tilde{\omega}'(\omega = 0) = \infty$, $\tilde{\omega}'(\omega = 1/6\beta) = 0$, and $\tilde{\omega}''(\omega) < 0$ for $\omega \in [0, 1/6\beta]$. Hence, a necessary and sufficient condition for a unique fixed point $\hat{\omega} \in (0, 1/6\beta)$ is $\tilde{\omega}(1/6\beta) < 1/6\beta$, or

$$(22)\quad \delta < \frac{4}{3\left[\sigma_v^{2f} + (\mathrm{E}V^f)^2\right] + 4} \equiv \delta_N.$$

This will clearly hold if δ and/or $[\sigma_v^{2f} + (\mathrm{E}V^f)^2]$ is sufficiently small. Thus, under this condition, the threat of interior Nash reversion generates a unique, most cooperative trade tax rate, with the properties given in Figure 3.

If instead (22) fails, then $\tilde{\omega}(1/6\beta) \geq 1/6\beta$. Since $\tau_c(V^f, \omega) = 0$ for all $\omega \geq 1/6\beta$, this case corresponds to free trade for every V^f. Taking these results together, we have now established a unique, most cooperative trade tax function, which can be expressed solely in terms of V^f and other exogenous parameters, such as δ, σ_V^{2f}, and EV^f.

The trade tax function is easily understood. When δ and $[\sigma_v^{2f}+(\mathrm{EV}^f)^2]$ are small and (22) holds, the threat of interior Nash reversion is unimpressive. Intuitively, since δ is small, future losses from defection are not weighted heavily, while a small $[\sigma_V^{2f}+(\mathrm{EV}^f)^2]$ implies that the reduced trade volume induced by the reversion is not expected to be large or variable and does not therefore represent a great loss in welfare. This case corresponds to Figure 3 with ω defined via the fixed point condition in terms of exogenous variables. If on the other hand δ and $[\sigma_V^{2f}+(\mathrm{EV}^f)^2]$ are large and violate (22), then the reversion threat is acute, and so free trade is sustainable for all V^f.

Note that (19) and (21) together imply that surges in trade volume will tend to lead to greater increases in protection the more "unusual" they are for the sector under consideration. That is, the level of protection sustainable in the cooperative equilibrium of the dynamic tariff game will depend on the realization of V^f and its mean EV^f on the one hand, and on the variance of V^f, σ_V^{2f}, on the other. If free trade is sustainable when $V^f=\mathrm{EV}^f$, then all else equal, a given increase in V^f above EV^f will be associated with a higher cooperative tariff the more "unusual" is the trade volume surge (the smaller is σ_V^{2f}).

To summarize, we have demonstrated that a high natural trade volume increases the incentive to defect. Protection is then needed to mitigate the volume of trade, thus sustaining the cooperative equilibrium. Higher values of δ and $[\sigma_v^{2f}+(\mathrm{EV}^f)^2]$ act to make punishment more severe, and therefore make cooperation easier facing any given free trade volume V^f.[14]

The basic results of our analysis also carry through if countries explicitly choose trade quantities (quotas) rather than prices (taxes).[15] In particular, suppose countries set import and export quotas, and then either give the chosen quantity of quota licenses to their firms or auction them off. As noted in Section III, the unique static Nash equilibrium of this game is autarky, which then characterizes the credible punishment for defection.[16] While as in the tariff game the current incentive to defect from the tacit cooperative quota is increasing in underlying trade volume, it is greater for a given trade volume in the quota game than in the tariff game, since defection in the quota game secures *all* the rents of protection for the defecting country.[17] Calculations similar to those above yield the following representation of the most cooperative quota rule $q^c(V^f,\omega)$:

$$(23)\quad q_c(V^f,\omega) = \begin{cases} V^f \text{ for } V^f\in[0,\bar{V}^f] \\ \dfrac{V^f}{2}-\dfrac{\sqrt{(V^f)^2-4\beta\omega}}{2} \\ \qquad \text{for } V^f\in(\bar{V}^f,1]. \end{cases}$$

Figure 4 summarizes (23). Here the cooperative quota level $q_c(V^f,\omega)$ is plotted against the underlying free trade volume V^f. For low to moderate levels of V^f, free trade is sustainable and is reflected in a movement along the 45° line. When V^f passes the threshold value of $\bar{V}^f=\sqrt{6\beta\omega}$, however, free trade is no longer sustainable as a cooperative equilibrium. Moreover, any

[14] The threat to revert to a static Nash equilibrium is not in this model renegotiation-proof, as defined by Joseph Farrell and Eric Maskin (1987). Our basic conclusion as to the relation between trade volume and protection is consistent with the potential for renegotiation, however. To construct punishment schemes that will not be renegotiated, specify asymmetric tariffs (off the equilibrium path) so that the country that did not cheat enjoys the punishment phase.

[15] Rotemberg and Saloner (1986b) develop a model to analyze the effect of quotas on the ability of firms to collude. We differ in focusing on tacit cooperation between governments when firms are competitive.

[16] Since there is no interior static equilibrium for the quota game, we focus on the autarky threat, which is the threat that generates the most cooperation. See also fn. 11.

[17] In terms of Figure 2, defection in the quota game captures the additional revenue rectangle *aklo* (*ajik*) for the importer (exporter), increasing the static gains from defection above that in the tariff game.

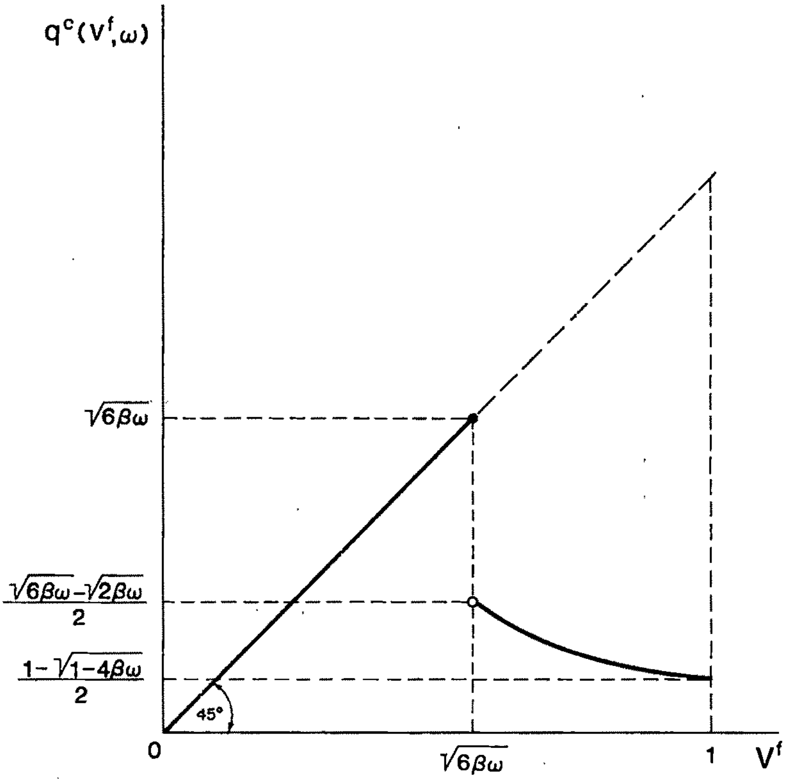


FIGURE 4

quota that restricts trade by less than the quota to which countries would optimally defect from free trade will only *increase* the static incentive to defect: since such a quota could not bind under defection, it would simply reduce the current welfare under cooperation for both countries but have no impact on current welfare under defection. For this reason, Figure 4 shows a discontinuity at $\bar{V}^f$, with higher free trade volumes leading to a discrete tightening of the trade restricting cooperative quota. At this lower cooperative quota, the optimal defection entails slightly tightening the quota so as to secure all cooperative quota rents. By choosing a sufficiently low cooperative quota, the countries reduce the size of cooperative quota rents and thus the incentive to defect.[18]

IV. Protection and the Trade Balance

In this section we explore how the relationship between trade *volume* and protection analyzed in the previous section translates into a relationship between trade *balance* and protection. While the model developed above relates levels of protection most directly to trade volume, there is a

[18] It is interesting to compare the levels of cooperation attainable in the tariffs and quota games. The autarky threat always provides the most cooperation, for a given policy instrument. Incentives to cheat are highest, however, when quotas are used, since all rents can be captured by defection. Thus, the threat of autarky reversion in the tariff game generally supports the most liberal trading environment. Whether quotas are preferred to tariffs with the interior reversion threat depends on the impatience of countries: the more the future is valued, the more likely it will be that tariffs with the threat of interior Nash reversion are inferior to quotas in supporting liberal trade. This comparison is further analyzed in Bagwell and Staiger (1988).

large informal empirical literature on trade imbalance as a determinant of protection, and it may be useful to explore the implications of our model in this regard.

The basic insight developed in the previous sections is that the incentive to defect unilaterally from cooperative tariff levels is highest when trade volume is highest: therefore, as the volume of trade expands, the level of protection rises in a cooperative equilibrium to mitigate the rising trade volume and hold the incentive to defect in check. When applied to a single sector as in the previous section, this implies that periods of greater than average trade volume will be associated with higher than average protection. We now show that when applied in the aggregate, this implies that the relationship between aggregate trade imbalance and the level of protection will depend on the sectoral makeup of the imbalance. A widening trade imbalance will be associated with greater levels of protection to the extent that the deficit (surplus) country's imports (exports) increase. However, a widening trade imbalance will be associated with lower levels of protection to the extent that the deficit (surplus) country's exports (imports) fall.

To show this, we consider the addition of a second sector to the model of the previous section and explore the relationship between the bilateral trade imbalance associated with trade between the home and foreign countries and their levels of protection. For simplicity, we suppose that the home and the foreign country trade only in these two products, and that there is a large rest of the world with which both countries also trade, so that the bilateral trade in these two goods is small relative to each country's total world trade. Thus, we stay within the partial equilibrium frameword of the previous sections and focus on the relationship between bilateral trade imbalance and protection.[19]

To keep things simple, suppose that the second product is identical to the first in every way except that while the first good is exported by the home country, the second good is exported by the foreign country. In particular, suppose foreign (home) output of good 1 is given by $e \in [0,1]$ $(2-e)$ and that the home (foreign) output of good 2 is given by $e \in [0,1]$ $(2-e)$. As before, e is determined independently in each period by the commonly known distribution $F(e)$. Under this setup, free trade volumes will be identical in the two sectors and the results of the previous section imply that the two goods will share identical levels of protection in the cooperative equilibrium for every realization of e (and thus V^f). Therefore, $\tau_c^1 = \tau_c^2 = \tau_c$ for all V^f, and τ_c is increasing in V^f. But the bilateral trade balance will be given by

$$(24)\quad \begin{aligned} TB &= [2-e-C_1]-[C_2-e] \\ &= 2-(C_1+C_2) \\ &= 2-(C_1+C_1^*) = 0, \end{aligned}$$

where the second-to-last equality follows from the symmetry of the setup which ensures that $C_2 = C_1^*$.

What we have shown is that observations on the trade balance are not generally sufficient to determine the path of protection: in the extreme case illustrated here, the trade balance is always zero, independent of the value of V^f, even though the cooperative level of protection will change with V^f, as was the case in the previous section. Hence, to predict the relationship between trade imbalance and protection, our model suggests that information on the sectoral makeup of the trade imbalance will be needed.[20]

[19] See Bergsten (1982) on the importance of Japan–U.S. bilateral trade imbalance in determining levels of protection between the two countries.

[20] A separate issue that arises with the introduction of more than one sector is the notion of multimarket contact developed by B. Douglas Bernheim and Michael Whinston (1987). The central idea is that multimarket contact allows the pooling of incentive constraints over markets: if slack exists in the incentive constraints of some markets, pooling may augment the degree of cooperation sustainable in other markets. In the special two-sector case considered above, there is no gain from pooling incentive constraints since the sectors are identical in all relevant ways. More gener-

V. Extensions

We discuss in this section several further extensions that can be introduced to the model. We begin by exploring the generality of the relationship between underlying free trade volume and the incentive for countries to defect that characterizes the model of the previous sections. The property that at least one country's incentive to defect from free trade rises during periods when the underlying free trade volume is high is crucial in generating the positive correlation between cooperative protection levels and trade volume depicted above. We now examine the generality of this relationship and provide support for a presumption in its favor.

Expressions (25) and (26) depict, respectively, the domestic (exporting) and foreign (importing) country's static incentive to defect from free trade for general export supply $(X(k, P))$ and import demand $(M(k^*, P^*))$ functions:

$$\begin{aligned}(25)\quad &\Omega(k,k^*,\tau_D,0)\\ &=\left[P^*(k,k^*,\tau_D,0)-P^f\right]\\ &\quad\times X(k,P(k,k^*,\tau_D,0))\\ &\quad-\int_{P(k,k^*,\tau_D,0)}^{P^f}\left[X(k,P)\right.\\ &\qquad\qquad\left.-X(k,P(k,k^*,\tau_D,0))\right]dP.\end{aligned}$$

$$\begin{aligned}(26)\quad &\Omega^*(k,k^*,0,\tau_D)\\ &=\left[P^f-P(k,k^*,0,\tau_D)\right]\\ &\quad M(k^*,P^*(k,k^*,0,\tau_D))\\ &\quad-\int_{P^f}^{P^*(k,k^*,0,\tau_D)}\left[M(k^*,P^*)\right.\\ &\qquad\qquad\left.-M(k^*,P^*(k,k^*,0,\tau_D))\right]dP^*.\end{aligned}$$

The first term in expressions (25) and (26) is the tariff revenue collected from one's trading partner. The second term in each expression is the efficiency loss in the surplus from trade associated with defection. The parameters k and k^* represent general positive shift parameters in the export supply and import demand functions, respectively. Thus, by definition $\partial X(k,P)/\partial k>0$ and $\partial M(k^*,P^*)/\partial k^*>0$ for all P and P^* and, provided $\partial X(k, P)/\partial P>0$ and $\partial M(k^*,P^*)/\partial P^*<0$, then $dV^f/dk>0$, and $dV^f/dk^*>0$.

We focus on the implications of a shift in the export supply or import demand function for the incentive to defect in the country where the shock originates; that is, we consider the signs of $d\Omega(\cdot)/dk$ and $d\Omega^*(\cdot)/dk^*$. We wish to establish a presumption that these two derivatives are positive, thereby ensuring that at least one country's incentive to defect from free trade rises whenever the underlying free trade volume rises. Using (25) and (26), direct calculation establishes the following:

$$(27)\quad \frac{d\Omega(\cdot)}{dk}>0\text{ iff }\frac{\partial X(k,P^f)}{\partial k}\left[\frac{P^f}{\eta_x^f+\eta_m^f}\right]>\int_{P(k,k^*,\tau_D,0)}^{P^f}\frac{\partial X(k,P)}{\partial k}dP,$$

$$(28)\quad \frac{d\Omega^*(\cdot)}{dk^*}>0\text{ iff }\frac{\partial M(k^*,P^f)}{\partial k^*}\left[\frac{P^f}{\eta_x^f+\eta_m^f}\right]>\int_{P^f}^{P^*(k,k^*,0,\tau_D)}\frac{\partial M(k^*,P^*)}{\partial k^*}dP^*,$$

where η_x^f and η_m^f are, respectively, the price elasticities of export supply and import demand (taken positively) evaluated at free trade. The right-hand side of (27) gives the additional efficiency loss from τ_D suffered by the defecting home country associated with the outward shift of its export supply function. Likewise, the right-hand side of (28) gives the additional efficiency loss from τ_D suffered by the defecting foreign country

ally, however, the ability of governments to pool incentive constraints across sectors is likely to undermine a strict sector-by-sector relationship between trade volume and protection. Instead, a surge in overall bilateral trade volume would lead to an overall rise in bilateral protection.

associated with the outward shift of its import demand function. The left-hand side of these two expressions captures the impact of the terms-of-trade changes associated with the shock on each country's incentive to defect: the less elastic are the export supply and import demand functions at free trade, the greater will be the (free trade) terms-of-trade loss for the country within which the shock originates, and the greater will be the corresponding gain from defection.

From (27) and (28) it is clear that a shock to the export supply or import demand function that takes the form of an increase in k or k^*, respectively, will increase the incentive to defect from free trade for the country within which the shock occures provided that $\eta_x^f + \eta_m^f$ are sufficiently small, that is, provided that the elasticities of export supply and import demand evaluated at free trade are sufficiently small. But in the context of a defection, this is likely to be the case, since the relevant elasticities are very short run in nature; that is, they reflect the time it takes for one's trading partner to detect a defection and respond. Thus, with a negligible (immediate) response of export supplies and import demands to a small price change from P^f, conditions (27) and (28) are likely to be satisfied, and at least one country's incentive to defect from free trade will rise whenever the underlying free trade volume rises. This supports the presumption in favor of an increasing relationship between free trade volume and the incentives of at least one country to defect from free trade, and suggests that the flavor (though not the transparency) of our results would be preserved in much more general settings.[21]

We turn next to the interpretation of the export tax. While it is true that import taxes are more commonly utilized than are export taxes, the latter are nonetheless observed, especially in the primary product markets of developing countries.[22] Still, it is interesting to ask whether our basic results would be preserved in a world in which the imposition of export taxes were politically infeasible. In this regard, it is possible to amend the two-sector model considered in Section IV with an *ad hoc* requirement that countries are unable to impose export taxes, perhaps because of political pressures. In this amended model, there is a unique static Nash equilibrium in which each country imposes the import tax $\tau_R(V^f,0) = 2V^f/3\beta$ on its import good. Cooperation in an infinite-horizon setting is then made possible with the threat to forever revert to the static equilibrium if a defection from a low tariff is ever observed. Our basic result readily extends to this model. During periods of high-volume trade, each country will have a large incentive to deviate to the tariff $2V^f/3\beta$ on its import good. This incentive is reduced only if the cooperative tariff is allowed to rise somewhat, so that each country is appeased in its import market. Thus, even if countries have no direct control over the volume of trade in their export sector, higher trade volume will continue to correspond to higher protection.

Finally, we have chosen to focus on governments that pursue protection for its terms-of-trade effects, and thus have considered a model in which full international cooperation would entail the maintenance of free trade in all periods. However, the basic insights we have developed are consistent with other government objectives. For example, suppose instead that import protection is used by governments to mitigate sudden drops in the real income of import-competing producers. Such an objective is

[21] We have also assumed throughout that e is i.i.d. through time. Rotemberg and Saloner (1986a) have made the analogous assumption for their model, but this approach has been criticized by John Haltiwanger and Joseph Harrington (1987). As applied to our model, the criticism is that if a high V^f today makes more likely a high V^f tomorrow, then the cost of defection today rises for both countries. The predictions of the i.i.d. model are reversed, however, only if δ is sufficiently large. Since the small δ case seems especially appropriate for elected governments, the possibility of correlated shocks is likely to be consistent with our conclusions for an interesting range of parameters.

[22] See, for example, *World Development Report 1986*, especially chap. 4, for a discussion of the export tax policies of developing countries.

in the spirit of Corden's (1974) Conservative Social Welfare Function. If trade volume into the importing country surges, either because of a "good" supply realization abroad or a "bad" supply realization at home, the importing country may respond with "special" protection even under full international cooperation. However, the fully cooperative response typically will trade off the desire to maintain the real income of import-competing producers at home not only with the interests of domestic consumers, but with the interests of the export sector abroad as well, and thus results in less complete maintenance of import-competing producer incomes than would be individually optimal. Thus, the workings of such a model would be qualitatively similar to the model we have studied here: unusual surges in trade volume would tend to be associated with unusually large static gains from defection to a high tariff for the government of the importing country, and an increase in the equilibrium level of protection (above its fully cooperative level but still below the noncooperative choice) would be required to avoid a complete breakdown in cooperation, that is, a tariff war.[23]

VI. Conclusions

We have attempted to develop a theory of managed trade that correlates periods of unusually high trade volume with increased protection. While the model we have chosen is special in a number of ways, the insights it generates appear to be much more general.

In particular, the notion that periods of unusually high trade volume present countries with an unusually strong incentive to defect from cooperative trading arrangements seems to be quite general and forms the heart of our analysis. Given this, it follows naturally that countries will attempt to manage the volume of trade with protective instruments that serve to dampen fluctuations in trade volume. Trade management can then be understood as an attempt by countries to maintain the self-enforcing nature of existing international cooperation.

Finally, in exploring the sustainable level of tacit cooperation among countries in a volatile environment, we have made no formal distinction between special protection devices that are provided for explicitly in the GATT and those that are not. The "safeguards" provisions of the GATT, whereby countries are given the right to raise protection in the event of unforseen developments, may to some extent represent an explicit institutional manifestation of our ideas.[24] Our analysis suggests a role for safeguard provisions when trade volume is unexpectedly high as a means of avoiding a reversion to noncooperative interaction among countries. In this light, the recent proliferation of safeguard "substitutes," for example, VERs and OMAs, may reflect the general inadequacy of the existing safeguards provisions to maintain the credibility of the rest of the GATT system.

[23] This particular interpretation is also suggestive of the way in which international agreements, even if constrained to be self-enforcing in nature, may help to limit the policy discretion of a government vis-à-vis its own private sector, thereby alleviating time-consistency problems of the kind studied in Staiger and Guido Tabellini (1987).

[24] Consistent with this interpretation is the observation of Kenneth Dam (1970): "One may conclude that the GATT escape clause is a useful safety valve for protectionist pressures and does not undercut in any serious way the advantages of the GATT tariff negotiating system. Insofar as the excape clause is a political 'prerequisite' to the membership in the GATT of certain contracting parties—most notably the U.S.—the argument in its favor is even stronger" (pp. 106–107). For a discussion of GATT safeguards, see also J. David Richardson (1988).

REFERENCES

Abreu, Dilip, "On the Theory of Infinitely Repeated Games with Discounting," *Econometrica*, March 1988, *56*, 383–96.

Bagwell, Kyle and Staiger, Robert W., "A Theory of Managed Trade," NBER Working Paper No. 2756, November 1988.

Benoit, Jean-Pierre and Krishna, Vijay, "Finitely Repeated Games," *Economet-*

rica, July 1985, *53*, 905–22.

Bergsten, C. Fred, "What To Do About U.S.–Japan Economic Conflict," *Foreign Affairs*, Summer 1982, *60*, 1059–75.

______ **and Cline, William R.,** "Trade Policy in the 1980s: An Overview," in William R. Cline, ed., *Trade Policy in the 1980s*, Cambridge, MA: MIT Press, 1983.

Bernheim, B. Douglas and Whinston, Michael D., "Multimarket Contact and Collusive Behavior," Harvard Discussion Paper No. 1317, May 1987.

Corden, W. Max, *Trade Policy and Economic Welfare*, Oxford: Clarendon Press, 1974.

Dam, Kenneth W., *The GATT: Law and International Economic Organization*, Chicago: University of Chicago Press, 1970.

Dixit, Avinash, "Strategic Aspects of Trade Policy," in Truman F. Bewley, ed., *Advances in Economic Theory: Fifth World Congress*, New York: Cambridge University Press, 1987, 329–62.

Farrell, Joseph and Maskin, Eric, "Renegotiation in Repeated Games," Harvard University Discussion Paper No. 1335, 1987.

Feenstra, Robert C. and Lewis, Tracy R., "Negotiated Trade Restrictions with Private Political Pressure," NBER Working Paper No. 2374, September 1987.

Friedman, James W., "Cooperative Equilibria in Finite Horizon Noncooperative Supergames," *Journal of Economic Theory*, April 1985, *35*, 390–98.

Haltiwanger, John and Harrington, Joseph E. Jr., "The Impact of Cyclical Demand Movements on Collusive Behavior," Working Paper No. 209, April 1988.

Jensen, Richard and Thursby, Marie, "Free Trade: Two Noncooperative Approaches," Ohio State University Working Paper, 1984.

______, "Tariffs with Private Information and Reputation," unpublished manuscript, 1989.

Johnson, Harry G., "Optimal Tariffs and Retaliation," *Review of Economic Studies*, 1953–54, *21*, 142–53.

Keenan, John and Riezman, Raymond, "Do Big Countries Win Tariff Wars?" *International Economic Review*, February 1988, *29*, 81–85.

Mayer, Wolfgang, "Theoretical Considerations on Negotiated Tariff Adjustments," *Oxford Economic Papers*, March 1981, *33*, 135–53.

Page, S. A. B., "The Management of International Trade," in Robin Major, ed., *Britain's Trade and Exchange Rate Policy*, London: Heinemann, 1979, 164–99.

Richardson, J. David, "Safeguards Issues in the Uruguay Round and Beyond," in Robert E. Baldwin and J. David Richardson, eds., *Issues in the Uruguay Round*, NBER Conference Report, 1988.

Riezman, Raymond, "Tariff Retaliation from a Strategic Viewpoint," *Southern Economic Journal*, January 1982, *48*, 583–93.

______, "Dynamic Tariffs with Asymmetric Information," University of Iowa unpublished manuscript, October 1987.

Rotemberg, Julio J. and Saloner, Garth, (1986a) "A Supergame-Theoretic Model of Price Wars During Boom," *American Economic Review*, June 1986, *76*, 390–407.

______, (1986b) "Quotas and the Stability of Implicit Collusion," MIT Discussion Paper No. 419, May 1986.

Staiger, Robert W. and Guido Tabellini, "Discretionary Trade Policy and Excessive Protection," *American Economic Review*, December 1987, *77*, 823–37.

Tower, Edward, "The Optimum Quota and Retaliation," *Review of Economic Studies*, October 1975, *42*, 623–30.

U.S. Trade Representative, *Annual Report of the President of the United States on the Trade Agreements Program*, 1983, Washington, April 1984.

World Development Report 1986, New York: Oxford University Press, 1986.

Comparative Advantage and Long-Run Growth

By Gene M. Grossman and Elhanan Helpman*

We construct a dynamic, two-country model of trade and growth in which endogenous technological progress results from the profit-maximizing behavior of entrepreneurs. We study the role that the external trading environment and that trade and industrial policies play in the determination of long-run growth rates. Cross-country differences in efficiency at R&D versus manufacturing (i.e., comparative advantage) bear importantly on the growth effects of economic structure and commerical policies. (JEL *411, 111, 621)*

What role do the external trading environment and commercial policy play in the determination of *long-run* economic performance? This central question of international economics has received surprisingly little attention in the theoretical literature over the years.

Previous research on trade and growth has adopted the neoclassical framework to focus on factor accumulation in the open economy. (See the surveys by Ronald Findlay, 1984, and Alasdair Smith, 1984.) This research largely neglects the effects of trade structure on rates of growth, however, addressing instead the reverse causation from growth and accumulation to the pattern of trade.[1] The direction that the research followed almost surely can be ascribed to the well-known property of the standard neoclassical growth model with diminishing returns to capital that (endogenous) growth in per capita income dissipates in the long run. For this reason, the familiar models that incorporate investment only in capital equipment seem ill-suited for analysis of long-run growth.

The available evidence collected since the seminal work of Robert Solow (1957) also leads one to look beyond capital accumulation for an explanation of growth. Exercises in growth accounting for a variety of countries generally find that increases in the capital to labor ratio account for considerably less than half of the last century's growth in per capita incomes.[2] Although econometric efforts to explain the residual have been somewhat disappointing (see, for example, Zvi Griliches, 1979), professional opinion and common sense continue to impute much of this residual to improvements in technology.[3] We share the view, expressed by Paul Romer (1986, 1990), that a full understanding of growth in the long run requires appreciation of the economic determinants of the accumulation of knowledge.

In this paper we draw on the pioneering work by Romer to construct a model that highlights the roles of scale economies and technological progress in the growth process. As in Romer's work, our model im-

*Grossman: Woodrow Wilson School, Princeton University, Princeton, NJ 08544. Helpman: Department of Economics, Tel Aviv University, Ramat Aviv, Israel. We are grateful to Avinash Dixit, Carl Shapiro, and Lars Svensson for their comments on an earlier draft, and to the National Science Foundation and the Alfred P. Sloan Foundation for financial support. Grossman also thanks the World Bank and Helpman thanks the International Monetary Fund for providing support and a stimulating environment during part of our work on this project. These organizations are not responsible for the views expressed herein.

[1] An exception is Max Corden (1971), who studies how the opening up of trade affects the speed of transition to the steady state in a two-factor neoclassical growth model with fixed savings propensities.

[2] See Angus Maddison (1987) for a careful, recent exercise in growth accounting.

[3] The benefits of education and experience undoubtedly contribute part of the explanation for the growth residual. See, for example, Robert Lucas (1988) and Gary Becker and Kevin Murphy (1990) for growth models that highlight the role of human capital accumulation as a source of growth.

plies an endogenous rate of long-run growth in per capita income, and we study its economic determinants. Our primary contribution lies in casting the growth process in a two-country setting. We provide, for the first time, a rigorous analysis linking long-run growth rates to trade policies and other international economic conditions. Moreover, we show that recognition of cross-country differences in economic structure impinge upon conclusions about the long-run effects of domestic shocks and policies.

Our model incorporates the essential insights from Romer (1990), although we introduce some differences in detail. The building blocks are an R&D sector that produces designs or blueprints for new products using primary resources and previously accumulated knowledge, an intermediate-goods sector consisting of oligopolistic producers of differentiated products, and a consumer-goods sector in each country that produces a country-specific final output using labor and intermediate inputs. As in Wilfred Ethier (1982), total factor productivity in final production increases when the number of available varieties of differentiated inputs grows. Thus, resources devoted to R&D contribute over time to productivity in the production of final goods, as well as to the stock of scientific and engineering knowledge.

The new elements in our analysis stem from the assumed presence of cross-country differences in the effectiveness with which primary resources can perform different activities, that is, *comparative advantage*. For simplicity we specify a one-primary-factor model, and allow the productivity of this factor in the three activities to vary internationally. Similar results could be derived from a multifactor model with interindustry differences in factor intensities (see Grossman and Helpman, 1989d, and Grossman, 1989). In any event, we find that many comparative dynamic results hinge on a comparison across countries of parameters reflecting efficiency in R&D relative to efficiency in manufacturing the goods that make use of the knowledge generated by R&D. The effects of policy in a single country, of accumulation of primary resources in a single country, and of a shift in world tastes toward the final output of one of the countries all depend upon the identity of the country in which the change originates in relation to the international pattern of comparative advantage.

We describe the economic setting in Section I below. Then, in Section II, we derive the dynamic equilibrium of the world economy and calculate two reduced-form equations that determine the steady-state growth rate. In Section III, we investigate the structural determinants of long-run growth. There, the implications for growth of variations in consumer preferences, primary-input coefficients in one or both countries, and stocks of available primary resources are considered. Section IV contains policy analysis. We study barriers and inducements to trade in consumer goods and subsidies to research and development. The analysis is extended in Section V to incorporate lags in the dissemination of knowledge and asymmetries in the speed of diffusion within and between countries. We use the extended model to reconsider the effects of trade policies on the steady-state rate of growth. Finally, Section VI provides a brief summary of our findings.

I. The Model

We study a world economy comprising two countries. Each country engages in three productive activities: the production of a final good, the production of a continuum of varieties of differentiated middle products (i.e., intermediate inputs), and research and development (R&D). A single primary factor is used in production, and is taken to be in fixed and constant supply in each country. Although we refer to this factor as "labor," we have in mind an aggregate of irreproducible resources that for any given level of technical know-how limits aggregate output.

At a point in time, output of final goods in country i is given by

$$Y_i = BA_i L_{Yi}^{1-\beta}\left[\int_0^n x_i(\omega)^\alpha \, d\omega\right]^{\beta/\alpha}, \qquad 0<\alpha,\beta<1,$$

where L_{Yi} represents employment in the final-goods sector, $x_i(\omega)$ denotes the input of middle product ω, A_i is a country-specific productivity parameter, and n is (the measure of) the number of varieties of middle products available at that time.[4] This production function exhibits constant returns to scale for given n, but an increase in the measure of varieties of middle products raises total factor productivity. This specification, which we borrow from Ethier (1982), captures the notion that an increasing degree of *specialization* generates technical efficiency gains. The economy's potential for augmenting the degree of specialization by developing new middle products implies the existence of dynamic scale economies at the industry level that are external to the individual final-good-producing firms.

Competition in the final-goods sectors ensures marginal-cost pricing. Hence, by appropriate choice of the constant B, producer prices satisfy

$$(1)\quad p_{Yi} = \left(\frac{w_i}{A_i}\right)^{1-\beta}\left[\int_0^n p_X(\omega)^{1-\varepsilon} d\omega\right]^{\beta/(1-\varepsilon)},$$

$$\varepsilon = \frac{1}{1-\alpha} > 1,$$

where w_i is the wage in country i and $p_X(\omega)$ is the price of variety ω. Final-good producers worldwide pay the same prices for (freely traded) middle products.

At every moment in time, the existing producers of middle products engage in oligopolistic price competition. The producer of a variety ω in country i chooses $p_X(\omega)$ to maximize profits,

$$\pi_i(\omega) = [p_X(\omega) - w_i a_{LXi}] \times \frac{p_X(\omega)^{-\varepsilon}}{\int_0^n p_X(\omega)^{1-\varepsilon} d\omega}\beta\Sigma_i p_{Yi}Y_i,$$

where a_{LXi} is the unit labor requirement for production of intermediates in country i. This expression for profits comprises the product of profits per unit (in square brackets) and derived demand for variety ω, where the latter incorporates the assumption that neither the prices of competing products nor the value of final production varies with $p_X(\omega)$. The first-order condition for a profit maximum implies the usual fixed-markup pricing rule,

$$(2)\qquad \alpha p_X(\omega) = w_i a_{LXi}.$$

It is clear from (2) that varieties originating from the same country bear the same price. Letting p_{Xi} represent the price of a variety produced in country i and n_i be the number of intermediate inputs produced there, equations (1) and (2) imply

$$(3)\quad p_{Yi} = \left(\frac{w_i}{A_i}\right)^{1-\beta}\left(\sum_j n_j p_{Xj}^{1-\varepsilon}\right)^{\beta/(1-\varepsilon)},$$

$$(4)\qquad \alpha p_{Xi} = w_i a_{LXi}.$$

With these prices, profits per firm can be expressed as

$$(5)\qquad \pi_i = (1-\alpha) p_{Xi} X_i / n_i,$$

where X_i is aggregate output of intermediates in country i (n_i times per-firm output) and is given by

$$(6)\qquad X_i = \frac{n_i p_{Xi}^{-\varepsilon}}{\sum_j n_j p_{Xj}^{1-\varepsilon}}\beta\left(\sum_j p_{Yj}Y_j\right).$$

As in our 1989c paper (see also Kenneth Judd, 1985), resources devoted to research generate "blueprints" that expand the measure of differentiated products. Research outlays are made by private, profit-maximizing entrepreneurs, who appropriate some of the benefits from a new technological innovation in the form of a stream of oligopoly profits. We assume that innovators receive indefinite patent protection, but that blueprints are not tradable, so that all profits must be derived from production in

[4] Here, and henceforth, we omit time arguments when no confusion is caused by doing so.

the country in which a middle product has been developed. Then, with free entry by entrepreneurs, if resources are devoted to R&D in country i at time t, the present value of future operating profits from producing there—discounted to time t—must equal the current cost of R&D. We denote R&D costs by $c_{ni}(t)$ and write the zero-profit condition as

$$\int_t^\infty e^{-[R(\tau)-R(t)]}\pi_i(\tau)\,d\tau = c_{ni}(t),$$

where $R(t)$ is the cumulative interest factor from time 0 to time t ($R(0)=1$). Differentiating this condition with respect to t, we find

$$\frac{\pi_i + \dot{c}_{ni}}{c_{ni}} = \dot{R}. \tag{7}$$

Equation (7) expresses a standard no-arbitrage condition. Recognizing that $c_{ni}(t)$ represents the value of an input-producing firm in country i at time t, (7) equates the instantaneous rate of return on shares in such a firm (the sum of dividends and capital gains) to the rate of interest.

We follow Romer (1990) in assuming that R&D generates a second output, which takes the form of a contribution to the stock of disembodied knowledge. Knowledge here includes all general scientific information, as well as some forms of engineering data with more widespread applicability, generated in the course of developing marketable products. Knowledge contributes to the productivity of further research efforts by reducing the amount of labor needed for an inventor to develop a new product. Due to the more general and non-patentable nature of this product of the R&D effort, appropriation of the resulting returns by the creator seems problematic. We assume to begin with that general knowledge disseminates immediately and costlessly throughout the world. This approximates a situation in which information spreads through technical journals, professional organizations, and interpersonal commercial contacts, and where literature, scientists, and business-people move freely across international borders (see Luigi Pasinetti, 1981, ch. 11). We relax this assumption by introducing lags in the dissemination of knowledge in Section V.

With these knowledge spillovers in mind, we specify our R&D technology as follows. If L_{ni} units of labor engage in research in country i, they generate a flow of new products $\dot{n}_i$ given by

$$\dot{n}_i = L_{ni}K/a_{Lni}, \tag{8}$$

where K is the current stock of knowledge and a_{Lni} is a country-specific productivity parameter. We take the stock of knowledge to be proportional to cumulative experience in R&D; that is, there are no diminishing returns to research in adding to scientific understanding. By choosing units for K so that the factor of proportionality is unity, we have $K = n$ and

$$\dot{K} = \sum_i L_{ni}K/a_{Lni}. \tag{9}$$

Since knowledge has been assumed to be a free input to each individual entrepreneur, the cost of product development in country i can be written as

$$c_{ni} = w_i a_{Lni}/n. \tag{10}$$

We turn now to the demand side of the model. Consumers worldwide share identical, homothetic preferences. They view the final goods produced in the two countries as imperfect substitutes. We represent preferences by a time-separable intertemporal utility function

$$U_t = \int_t^\infty e^{-\rho(\tau-t)} \times \log u[y_1(\tau), y_2(\tau)]\,d\tau, \tag{11}$$

where ρ is the subjective discount rate and $y_i(\tau)$ is consumption of final goods from country i in period τ. The instantaneous sub-utility function $u(\cdot)$ is nondecreasing, strictly quasi-concave, and positively linearly homogeneous.

A typical consumer maximizes (11) subject to an intertemporal budget constraint. With $u(\cdot)$ linearly homogeneous, this optimization problem can be solved in two stages. First, the consumer maximizes static utility for a given level of expenditure at time τ, $E(\tau)$. The solution to this subproblem generates an indirect utility function, $v[p_{Y1}(\tau), p_{Y2}(\tau)]E(\tau)$, where p_{Yi} is the price of Y_i. In the absence of barriers to trade in final goods, these prices are common to consumers in the two countries. The second-stage problem involves choosing the time pattern of expenditures to maximize

$$(12)\quad V_t = \int_t^{\infty} e^{-\rho(\tau-t)}\{\log v[p_{Y1}(\tau), p_{Y2}(\tau)] + \log E(\tau)\}\, d\tau$$

subject to

$$(13)\quad \int_t^{\infty} e^{-[R(\tau)-R(t)]} E(\tau) \le \int_t^{\infty} e^{-[R(\tau)-R(t)]} w(\tau) L\, d\tau + Z(t),$$

where $w(t)$ is the consumer's wage rate at time t, L is his labor supply, and $Z(t)$ is the value of his time t asset holdings. The interest factor in (13) is common to all individuals as a result of trade on the integrated world capital market, but the wage rate varies by country.

From the first-order conditions to this problem, we find that the optimal path for expenditure obeys

$$(14)\quad \frac{\dot{E}}{E} = \dot{R} - \rho.$$

Savings are used to accumulate either ownership claims in input-producing firms or riskless bonds issued by these same firms. Arbitrage ensures that the rates of return on these two assets are equal, and in equilibrium consumers are indifferent as to the composition of their portfolios.

II. Equilibrium Dynamics

During the course of the development of our model in the previous section, we provided some of the equilibrium conditions. For example, we derived pricing equations for goods and a no-arbitrage condition relating equilibrium asset returns. In this section we complete the list of equilibrium requirements by adding conditions that stipulate market clearing in factor and final-goods markets. We then derive and discuss a reduced-form system that describes equilibrium dynamics.

Static equilibrium in the markets for the two final goods implies

$$(15)\quad p_{Yi}Y_i = s_i E,$$

where $s_i(p_{Y1}, p_{Y2})$ is the share of world spending allocated to Y_i and E is world spending on consumer goods. The share function is, of course, homogeneous of degree zero. We establish below that relative commodity prices are constant in the vicinity of a steady state with active R&D sectors in both countries. For this reason, we take s_i to be constant in our subsequent analysis, and omit its functional dependence on relative prices.

The labor-market clearing conditions equate labor supply and labor demand in each country. Using (3) and Shephard's lemma, we see that final-goods producers demand $(1-\beta)p_{Yi}Y_i / w_i$ workers. The demand for labor by middle-products manufacturers is $a_{LXi}X_i$, while (8) and the fact that $K = n$ imply demand for labor by product developers of $(a_{Lni}/n)\dot{n}_i$. Hence,

$$(16)\quad (a_{Lni}/n)\dot{n}_i + a_{LXi}X_i + (1-\beta)p_{Yi}Y_i / w_i = L_i,$$

where L_i is the labor force available in country i.

Since we neglect here the monetary determinants of the price level, we may choose freely a time pattern for one nominal variable. It proves convenient to specify the numeraire as follows:

$$p_{X1} = n(a_{LX1}/a_{Ln1})^{1/\varepsilon}. \tag{17a}$$

We show in Appendix A of our 1989a working paper that, with this normalization, a necessary condition for convergence to a steady state with positive R&D in both countries (i.e., nonspecialization) is

$$p_{X2} = n(a_{LX2}/a_{Ln2})^{1/\varepsilon}. \tag{17b}$$

Together, (17a) and (17b) imply that relative prices of middle products are constant along the convergent path, which further implies with (4) the constancy of relative wages, and with (3) the constancy of relative prices of final goods. This last fact justifies our treatment of expenditure shares as fixed.

Let $g(t)$ denote the rate of growth of the number of products and the stock of knowledge; that is, $g \equiv \dot{n}/n = \dot{K}/K$. Then from (17) and (4) we see that prices of intermediates and wages grow at rate g, while from (10), product development costs are constant. Equations (5)–(7), (10), (15), and (17) imply

$$X_i = \frac{n_i b_i^{1/\alpha}}{\Sigma_j n_j b_j}\frac{\beta E}{n} \tag{18}$$

and

$$\dot{R} = \frac{1}{\varepsilon - 1}\frac{\beta E}{\Sigma_j n_j b_j}, \tag{19}$$

where $b_i \equiv (a_{Lni}/a_{LXi})^{\alpha}$. The coefficients b_i will serve as our measures of *comparative advantage*. Country 1 enjoys comparative advantage in conducting R&D if and only if $b_1 < b_2$.

Since wages grow at the same rate as n, it proves convenient to define $e \equiv E/n$. Letting $\sigma_i \equiv n_i/n$ be the share of products manufactured in country i and noting that $g = \Sigma_i \dot{n}_i/n$, (16), (15), (17), (4), and (18) imply

$$g = H - \frac{\beta e}{\sigma} - \frac{1-\beta}{\alpha}se, \tag{20}$$

where we have defined $H \equiv \Sigma_i L_i/a_{Lni}$, the total *effective* labor force, $\sigma \equiv \Sigma_i \sigma_i b_i$, a weighted average of the comparative advantage parameters with product shares as weights, and $s \equiv \Sigma_i s_i/b_i$. Observe that the parameter σ, which provides a useful summary of the static intersectoral resource allocation, grows (shrinks) over time if and only if the growth rate of the number of differentiated middle products in the country with comparative *disadvantage* in R&D exceeds (falls short of) that of the other country.

We are now prepared to derive two equations that describe the dynamic evolution of the world economy. From the definition of e, we have $\dot{e}/e = \dot{E}/E - g$, or, substituting (14), (19), and (20),

$$\frac{\dot{e}}{e} = \frac{\beta e}{\alpha\sigma} + \frac{1-\beta}{\alpha}se - H - \rho. \tag{21}$$

Hence, the rate of increase of spending per middle product is larger the greater is spending per product and the smaller is the share of the country with comparative disadvantage in R&D in the total number of varieties.

Now, from the definition of the product shares σ_i, their rates of change are given by $\dot{\sigma}_i/\sigma_i = \dot{n}_i/n_i - \dot{n}/n$. Using (16) together with (17), (4), (18), and (20), we obtain

$$\dot{\sigma}_i = h_i - \frac{s_i}{b_i}\frac{1-\beta}{\alpha}e - \sigma_i\left(H - \frac{1-\beta}{\alpha}se\right), \tag{22}$$

where $h_i \equiv L_i/a_{Lni}$ is effective labor in country i, and $\Sigma_i h_i = H$. Since the evolution of the two product shares are related by $\Sigma_i \dot{\sigma}_i = 0$, we can replace (22) by a single differential equation in σ. Making use of the fact that $\dot{\sigma} = \Sigma_i \dot{\sigma}_i b_i$, we find

$$\dot{\sigma} = h - \frac{1-\beta}{\alpha}e - \sigma\left(H - \frac{1-\beta}{\alpha}se\right), \tag{23}$$

where $h \equiv \Sigma_i h_i b_i$.

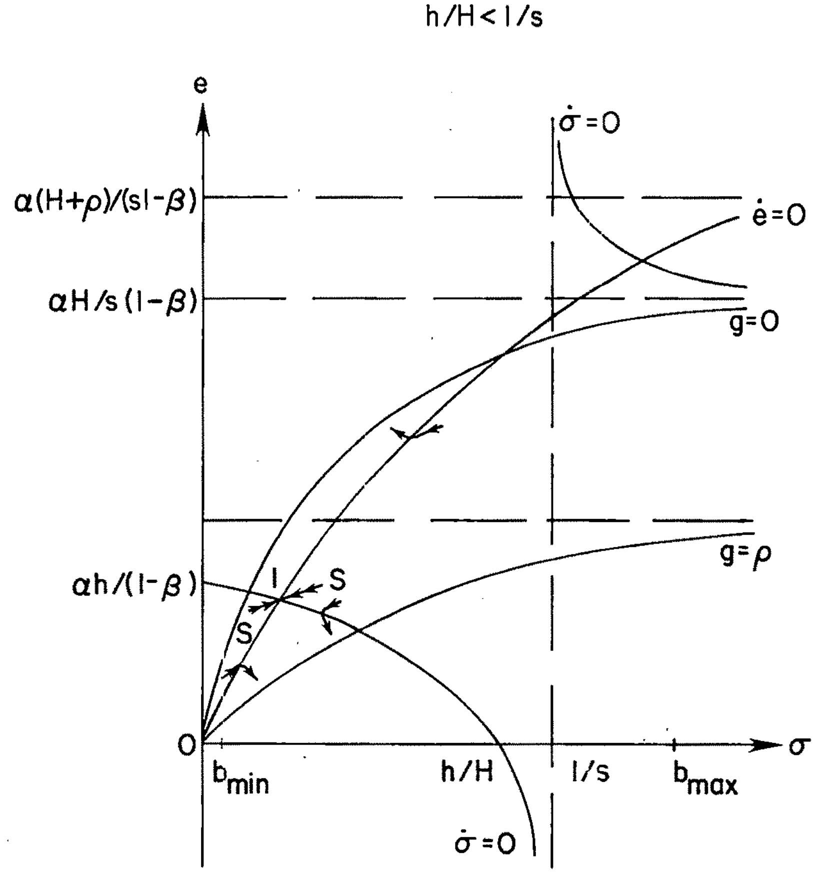


FIGURE 1

Equations (21) and (23) constitute an autonomous system of differential equations in e and σ. The solution to this system, together with (13), (20), and the definition of σ, provide a complete description of the evolution of spending and the number of products in each country. From these, the paths for outputs, employments and final-goods prices are easily derived. Thus, we shall use this two-equation system to analyze equilibrium dynamics.[5]

In Figures 1 and 2 we depict the stationary points for e. We draw the $\dot{e}=0$ locus as increasing and concave (see (21)). To understand the positive slope of this curve, observe from (19) that the interest rate can be expressed as

$$(24)\qquad \dot{R}=\beta e/\sigma(\varepsilon-1).$$

Thus, an increase in spending per product increases the interest rate and (from (20)) reduces the rate of growth of n (the former because the profitability of R&D rises with

[5] A restriction that we have not explicitly taken into account to this point is that product development must be nonnegative in each country. This condition certainly is satisfied in the neighborhood of a steady-state equilibrium with positive growth, but it need not hold all along the convergent path to such an equilibrium. Thus, strictly speaking, the equilibrium dynamics that we describe below apply for sure only in the vicinity of a steady state.

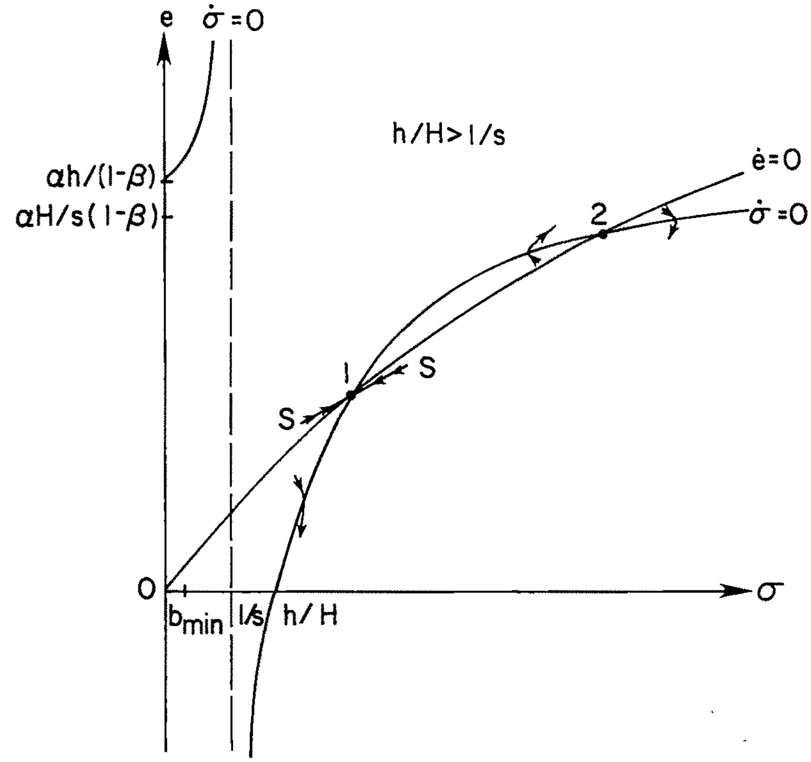


FIGURE 2

derived demand, the latter because more spending means less savings and hence less investment). Since an increase in the interest rate raises the rate of growth of nominal spending, and the rate of growth of e is just the difference between the rates of growth of E and n, it follows that an increase in e raises the growth rate of e. To compensate for this acceleration in spending per product, if e is to be stationary, σ must rise. An increase in σ lowers the interest rate and raises the rate of growth of n, thereby reducing the rate of growth of e.

Next, we distinguish two subcases depending on the relative sizes of h/H and $1/s$.[6] It can be shown that $h/H > 1/s$ if and only if $(b_2 - b_1)(h_2 b_2/s_2 - h_1 b_1/s_1) > 0$. Therefore, the case $h/H > 1/s$, depicted in Figure 2, applies, for example, if the shares of the two countries' final outputs are in proportion to their relative *effective* labor forces; that is, $h_1/s_1 = h_2/s_2$. But a bias in size relative to budget share of final output can reverse the inequality and hence the relationship between h/H and $1/s$. This gives us the case $h/H < 1/s$, shown in Figure 1, with which we begin the discussion.

In general, both $1/s$ and h/H must lie between $b_{\min}$ and $b_{\max}$. The $\dot{\sigma} = 0$ curve in Figure 1 is everywhere downward sloping, crosses the horizontal axis at h/H, and is discontinuous at $\sigma = 1/s$. The slope of the curve is understood as follows. For $\dot{\sigma} = 0$, we must have $\dot{\sigma}_1 = 0$, which requires that the resources available for R&D in each country be just sufficient to preserve the country's *share* in the world's number of

[6] A borderline case arises when $b_1 = b_2$; that is, when comparative advantage is absent. Then $h/H = 1/s$, and $\sigma = b$ always. Convergence to the steady-state level of e is immediate in this case; see Grossman and Helpman (1989a).

varieties. Consider country 1 and suppose for concreteness that this country has comparative advantage in R&D. Then an increase in σ lowers σ_1, thereby reducing the resources needed for production of middle products. The fall in σ_1 also reduces the amount of R&D country 1 must perform to preserve its share in the number of products. *Ceteris paribus*, σ_1 would tend to rise. An increase in e, on the other hand, diverts resources away from R&D to production of middle and final products in country 1. But it also causes the world's rate of product growth to fall, thereby diminishing the amount of R&D country 1 must undertake to maintain its share of middle products. The relative magnitudes of these two effects depend upon country 1's relative size, and on the share of its final product in aggregate spending. In the case under consideration, the second effect dominates, and so the $\dot{\sigma} = 0$ curve slopes downward.

In this case, there exists a unique steady state shown as point 1 in the figure. For initial values of σ not too different from that at point 1, a unique trajectory (saddle-path) converges to the steady state. This trajectory, labeled SS, fulfills all equilibrium requirements and satisfies the intertemporal budget constraint with equality. Along this trajectory (in the vicinity of the steady state), the interest rate and profit rate are declining (see (24)) and nominal expenditure E is rising. If the country with comparative advantage initially has a share of products that is smaller (larger) than its steady-state share, expenditure rises more slowly (rapidly) than the number of products.

The case depicted in Figure 2 arises when $h/H > 1/s$. Then the $\dot{\sigma} = 0$ schedule slopes upward. If the curve intersects the $\dot{e} = 0$ locus in the positive orthant at all, it must intersect it twice, as at points 1 and 2.[7] The lower point (point 1) represents the steady state with the higher rate of growth (growth rates increase as we move down along the $\dot{e} = 0$ schedule, as we demonstrate below) and indeed the growth rate corresponding to point 2 may be negative. More importantly, as we show in Appendix B of our 1989a working paper, the equilibrium at point 1 exhibits saddle-path stability, whereas that at point 2 is locally unstable.[8] To the right of point 1, the saddle-path leading to that point remains trapped in the area bounded by the $\dot{e} = 0$ locus and the line segment joining points 1 and 2, and is everywhere upward sloping. Thus, the qualitative properties of the dynamic trajectory that leads to a stable, positive-growth equilibrium in Figure 2 mimic those of the stable saddle-path in Figure 1.

For the remainder of this paper we shall restrict our discussion to stable steady-state equilibria with positive rates of growth. That is, we focus our attention on equilibria such as those at the points labeled 1 in Figures 1 and 2. In the steady state there occurs intraindustry trade in middle products and interindustry trade in consumer goods, with the long-run pattern of trade determined by comparative advantage, productivities in the two final-goods industries, and consumer preferences.

III. Determinants of Long-Run Growth

Our model generates an endogenous rate of long-run growth. We now are prepared to explore how economic structure and economic policy affect this growth rate. In this section, we derive the implications of sec-

[7] The geometry supports this claim, once we recognize that the $\dot{\sigma} = 0$ curve asymptotes to the horizontal line at $\alpha H/s(1-\beta)$, whereas the $\dot{e} = 0$ curve asymptotes to the horizontal line at $\alpha(H+\rho)/s(1-\beta)$. The algebra provides confirmation, as simple manipulation reveals that the steady-state growth rate solves a quadratic equation.

[8] We strongly suspect, however, that whenever there exist two positive-growth, steady-state equilibria in the admissible range, there also exists a third (saddle-path stable) steady-state equilibrium with zero growth. We have established the existence of such an equilibrium for some parameter values, but so far have been unable to construct a general existence proof. Since the equilibrium at point 1 in Figure 2 can only be reached if the initial value of σ is less than that at point 2, we suspect that initial values of σ in excess of that at point 2 (and perhaps only these) imply convergence to a steady state with zero growth.

toral productivity levels, country sizes, and demand composition for the steady-state growth rate. The influence of trade policies and of subsidies to R&D are treated in the next section.

We derive the long-run values of e and σ by setting $\dot{e}$ and $\dot{\sigma}$ to zero in (21) and (23). The steady-state magnitudes, $\bar{e}$ and $\bar{\sigma}$, solve

$$\frac{\beta\bar{e}}{\alpha\bar{\sigma}}+\frac{1-\beta}{\alpha}s\bar{e}=H+\rho; \tag{25}$$

$$\frac{1-\beta}{\alpha}\bar{e}(1-\bar{\sigma}s)+\bar{\sigma}H=h. \tag{26}$$

Whenever $1/s>h/H$, these equations provide at most one solution for $(\bar{e},\bar{\sigma})$ consistent with $\bar{g}>0$. When $1/s<h/H$, there may be two such solutions, in which case we select the stable equilibrium, that is, the one with the smaller values for $\bar{e}$ and $\bar{\sigma}$. Stability implies, in this latter case, that the $\dot{\sigma}=0$ curve intersects the $\dot{e}=0$ curve from below (see Appendix B of our 1989a working paper). We make use of this condition, namely,

$$\frac{\beta\bar{e}}{(H+\rho)\bar{\sigma}^2}>\frac{\alpha H-(1-\beta)s\bar{e}}{\bar{\sigma}H-h} \qquad \text{for} \quad 1/s<h/H, \tag{27}$$

in signing the comparative-dynamics derivatives that follow.

The growth rate of the number of varieties in the steady-state equilibrium can be derived from the solution to (25) and (26), together with (20). From this, we can easily calculate the growth rate of output. In the steady state, nominal expenditure grows at rate $\bar{g}$, while (3) implies that p_{Yi} grows at rate $[1-\beta/(1-\varepsilon)]\bar{g}$. From these facts and (15), we deduce that final output grows at rate $\beta\bar{g}/(1-\varepsilon)$.

It is worth noting at this point that the steady-state equations (25) and (26), as well as the equation for $\bar{g}$, do not rely on our assumption of perfect capital mobility. In the absence of capital mobility, the steady state would be the same as long as consumers worldwide share identical preferences (and therefore common subjective discount rates).[9]

It is instructive to begin the discussion with the case in which neither country exhibits comparative advantage in conducting R&D; that is, $b_1=b_2=b$. This case has $h=bH$ and $s=1/b$. Then (25) and (26) provide a unique solution for $\bar{e}$ and $\bar{\sigma}$, which upon substitution into (20) yields the long-run growth rate

$$\bar{g}=\frac{\beta(H+\rho)}{\varepsilon}-\rho. \tag{28}$$

This equilibrium growth rate shares much in common with that derived by Romer (1990) for a closed economy. In particular, the growth rate rises with effective labor H and declines with the subjective discount rate. Our measure of effective labor adjusts raw labor for productivity in R&D (recall that $H=\Sigma_i L_i/a_{Lni}$), so greater effectiveness in research in either country, as well as a larger world labor force, necessarily means faster growth. Long-run growth does not, however, depend upon coefficients that determine absolute productivity in the intermediate or final goods sectors (such as A_i or a_{LXi}). Nor do properties of the instantaneous utility function $u(\cdot)$, including the product composition of final demand, play any role in the determination of $\bar{g}$. As we shall see presently, all these features (except for the absence of an effect of A_i on $\bar{g}$) are special to a world without any comparative advantage.

Consider next the case with $1/s>h/H$. The curves $\dot{e}=0$ and $\dot{\sigma}=0$ in Figure 3 describe the initial situation, with a unique initial steady state at point 1. Now suppose

[9] The cases of perfect and imperfect capital mobility do differ in their implications for the steady-state share of each country in aggregate spending E. However, as should be clear from (25) and (26), the cross-country composition of E does not matter for the issues taken up in the present section.

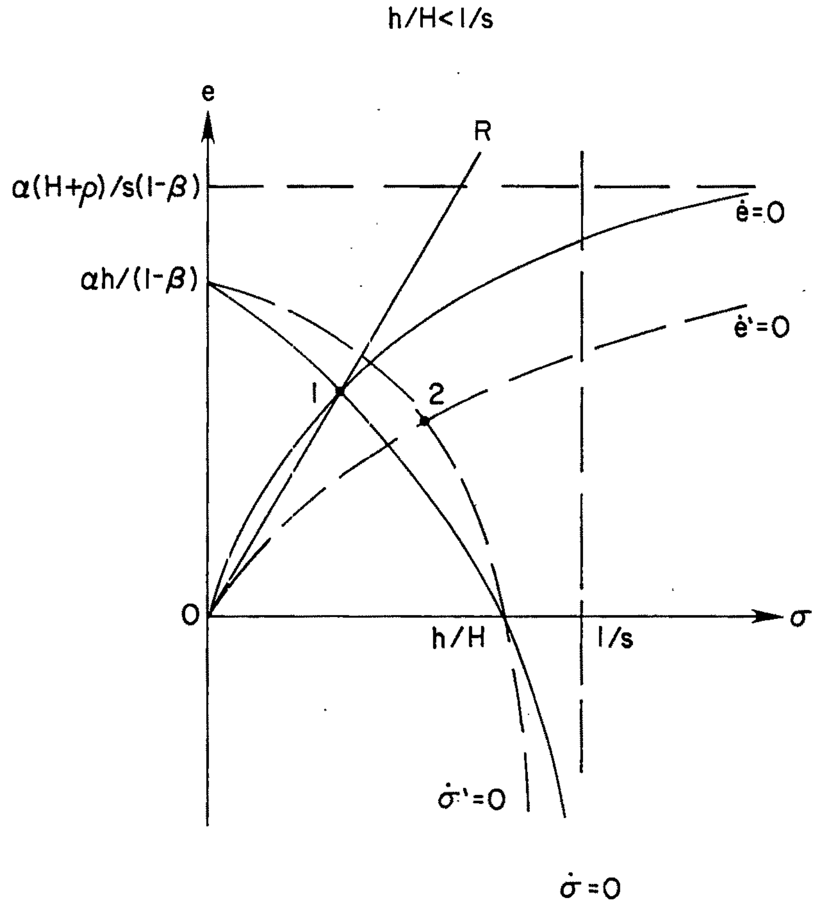


FIGURE 3

that preferences change so that s increases. This corresponds to a shift in tastes in favor of the final good produced by the country with comparative advantage in performing R&D. From (25), we see that the $\dot{e}=0$ curve shifts down, say to $\dot{e}'=0$ in the figure. Equation (26) implies that the $\dot{\sigma}=0$ schedule shifts out (in the positive orthant) to $\dot{\sigma}'=0$. The new steady state occurs at a point such as 2. But observe that all points on $\dot{e}'=0$ to the right of its intersection with ray OR are characterized by slower steady-state growth than at point 1. This claim follows from (20) and (25), whence

$$\bar{g}=\frac{\beta\bar{e}}{(\varepsilon-1)\bar{\sigma}}-\rho. \tag{29}$$

Since the intersection of $\dot{\sigma}'=0$ and $\dot{e}'=0$ necessarily lies to the right of the intersection of the latter curve with OR, we have established that an increase in s reduces steady-state growth.

When tastes shift unexpectedly toward the final good of the country with comparative advantage in R&D, resources there must be reallocated to satisfy the relatively higher consumer demand. A process begins whereby labor there shifts out of R&D and the manufacture of middle products. Products accumulate more slowly in this country than in the other, and over time its share of middle products falls (i.e., σ rises). Output per middle product changes by the same proportion in both countries (see (18)). So, in the new steady state, the country with comparative disadvantage in R&D is responsible for a relatively larger share of the world's innovation, with adverse consequences for the common steady-state growth rate. Of course, the opposite conclusion applies when s falls. Moreover, the same results obtain at stable equilibrium points when $1/s < h/H$ (see our 1989a working paper). We have thus proven the following:

PROPOSITION 1: *Stronger relative demand for the final good of the country with comparative advantage in R&D lowers the long-run share of that country in the number of middle products and slows long-run growth of the world economy. In the absence of comparative advantage in R&D, the long-run growth rate is independent of the relative demand for final goods.*

Next we consider the dependence of growth on the sizes of the effective labor forces. Effective labor may grow without affecting cross-country comparative advantage either because the stock of irreproducible resources expands, or because the productivity of labor in all uses (or in R&D and intermediate-good production) rises equiproportionately. In the first experiment, suppose that both countries experience equiproportionate, once-and-for-all increases in the sizes of their effective labor forces. We have already seen that this change would augment world growth in the absence of comparative advantage. In our 1989a working paper we prove that the same result carries over to a world with intercountry differences in relative productivities at R&D and manufacturing. We establish the following there:

PROPOSITION 2: *An equiproportionate, once-and-for-all increase in the effective labor forces of both countries accelerates long-run growth.*

Greater resources generate faster growth in our model, as in Romer (1990), essentially because dynamic scale economies characterize long-run production.

We investigate next the effects of an increase in the effective labor force of a single country. Conceptually, it proves convenient to decompose this change into two elements. First, we increase h and H by the same percentage amount. This percentage is chosen to equal the product of the share of the expanding country in the world's effective stock of labor times the percentage increase in effective labor force that the expanding country actually experiences. This accounts for the total percentage change in H when H_i changes. Then we must adjust h with H fixed to arrive at the appropriate change in h.

As an intermediate step, let us consider the effects of an increase in h alone. This corresponds to an increase in the effective labor of the country with comparative disadvantage in R&D, and a decrease in the effective labor of the other, so that the sum remains constant. This imaginary reshuffling of the world's resources shifts the $\dot{\sigma} = 0$ schedule upward when $1/s > h/H$, and downward otherwise. In either case, the $\dot{e} = 0$ curve is unaffected and the new steady-state point lies on this curve to the right of the original point. Noting (29), this proves the following:

LEMMA 1: *A reallocation of resources between countries that maintains a constant world stock of effective labor raises the long-run growth rate and increases the long-run product share of the relatively R&D efficient country if and only if the share of this country in effective labor increases.*

When the effective labor force of only country 1 (say) increases, h rises by proportionately more or less than H, according to whether country 1 has comparative disadvantage or advantage in R&D. If country 1 has comparative advantage in R&D, then both the uniform increase in H and h and the adjustment (lowering) of h, that together comprise the effect of an increase in H_1, serve to accelerate world growth. But if country 1 has comparative disadvantage in R&D, the two effects work in opposite directions. The increase in resources, by Proposition 2, speeds growth; but the reallocation of given resources, by Lemma 1, slows growth. The net effect is ambiguous, as the numerical examples that we present in Appendix D of our 1989a working paper serve to demonstrate. We have established the following:

PROPOSITION 3: *The long-run growth rate is higher the larger is the effective labor force of the country with comparative advantage in R&D. A larger effective labor force in the country with comparative disadvantage in R&D may be associated with faster or slower growth, depending upon the extent of productivity differences. In the absence of comparative advantage, long-run growth is faster the larger is the effective labor force of either country.*

These results emphasize the novel features of growth in a world with distinct countries and intercountry differences in relative productivities. They also suggest that findings reported by Paul Krugman (1990) may be somewhat special. A country need not enjoy faster growth by joining the integrated world economy, if the country enjoys substantial comparative advantage in R&D. Moreover, growth in resources or improvements in the productivity of existing resources do not guarantee faster long-run growth in a world equilibrium with free trade. If resources expand or become more efficient in the country with comparative disadvantage in R&D, then the resulting intersectoral reallocation of resources worldwide might slow innovation and growth everywhere.

IV. Economic Policy

In this section we discuss the effects of tariffs, export subsidies, and R&D subsidies on long-run growth. In order to do so, it is necessary for us to introduce the relevant policy parameters into the equations that describe instantaneous and steady-state equilibrium. To avoid repetition of the detailed arguments presented in Section I, we present here only the necessary modifications of the model and then explain their implications for the steady-state conditions. We restrict attention to small taxes and subsidies; this restriction facilitates exposition, as the channels through which economic policies affect long-run growth can be seen more clearly. We confine our analysis of trade policies to those that impede or encourage trade in final goods.

The introduction of taxes and subsidies to the model necessitates consideration of the government's budget. As usual, we assume that the government collects and redistributes net revenue by lump-sum taxes and subsidies. In a static framework, this specification suffices to determine completely the government's budgetary policy. But in a dynamic framework the budget need not balance period by period, so budgetary policy in general must specify the intertemporal pattern of lump-sum collections and transfers. However, with perfectly foresighted and infinitely lived agents, our model exhibits the Barro-Ricardo neutrality property. Hence, we need not concern ourselves with the intertemporal structure of budget deficits so long as the present value of the government's net cash flow equals zero.

The presence of the aforementioned policies modifies the decision problem for consumers in country 1 in two ways. First, we replace the price of good i in (12) by $T_i p_{Yi}$, where $T_1 = 1$. With this formulation, p_{Yi} remains the producer price of final good i, $T_2 > 1$ represents a tariff in country 1 on imports of consumer goods, and $T_2 < 1$ represents a subsidy by country 2 on exports of final output.[10] Second, we add the present value of net taxes to the right-hand side of (13) as a lump-sum addition to consumer wealth. The amount of this collection or redistribution will differ across countries according to their policies.

These modifications do not affect (14), which continues to describe the optimal intertemporal pattern of expenditures for consumers worldwide as a function of the pattern of equilibrium interest rates. In a steady state with $\dot{e} = 0$, (14) reduces to

$$\dot{R} = \bar{g} + \rho. \tag{30}$$

[10] The effects of a country 2 import tariff and a country 1 export subsidy can be derived symmetrically, so we neglect these policies here and leave the maximand for consumers in country 2 as before.

Notice that (30) implies that in any steady state in which countries grow at the same rate, long-run equalization of interest rates obtains. This property of our model holds irrespective of the presence or absence of international capital mobility and the presence or absence of tariffs or export subsidies on final goods and subsidies to research and development.

Turning to the production side, our policies do not alter equations (3)–(8) describing pricing and output relationships in the intermediate and final-goods sectors and the technology for knowledge creation. However, R&D subsidies do change the private cost of R&D. We replace (10) by

$$c_{ni} = w_i a_{Lni}/nS_i, \tag{10'}$$

where $S_i > 1$ represents subsidization of research costs in country i. It proves convenient to redefine our numeraire to normalize for the effect of the R&D subsidy on the price of intermediate inputs in country 1. Our new normalization dictates a modified equation for the price of intermediates produced in country 2 as well. Together, these relationships, which replace (17a) and (17b), can be written as

$$p_{Xi} = n(S_i a_{LXi}/a_{Lni})^{1/\varepsilon}. \tag{17'}$$

As for the market-clearing conditions, the factor-markets equation (16) is not affected, but we must replace (15) by

$$p_{Yi}Y_i = \frac{s_{i1}E_1}{T_i} + s_{i2}E_2, \tag{15'}$$

where E_i denotes aggregate spending by consumers in country i, and the shares of spending devoted to good i by residents of country 1 and country 2 are $s_{i1} = s_i(p_{Y1}, p_{Y2}T_2)$ and $s_{i2} = s_i(p_{Y1}, p_{Y2})$, respectively. Although import tariffs and export subsidies on final goods do not affect steady-state producer prices of final output in our model,[11] the direct response of spending shares in country 1 to changes in trade policy must now be taken into account. Moreover, R&D subsidies, if introduced at different rates in the two countries, will affect the steady-state value of p_{Y1}/p_{Y2}, and may therefore influence the long-run spending shares in both countries.

This completes the necessary modifications of the equilibrium relationships. We can now use the extended model to derive the equations describing steady-state equilibrium in the presence of policy intervention. In a steady state, employment in the R&D sector is given by $a_{Lni}\dot{n}_i/n = a_{Lni}\bar{g}\bar{\sigma}_i$. Making use of (4), (5), (7), (30), (10′), and (17′) (which together imply $\dot{c}_{ni} = 0$ in a steady state), we find employment in the manufacture of middle products equal to $a_{Lni}\bar{\sigma}_i(\bar{g} + \rho)(\varepsilon - 1)/S_i$. Substitution of these terms into (16) yields the steady-state labor market–clearing condition,

$$\bar{g}\bar{\sigma}_i + \frac{(\varepsilon-1)(\bar{g}+\rho)}{S_i}\bar{\sigma}_i + \frac{1-\beta}{\alpha b_i S_i^{1/\varepsilon}}\bar{q}_i = h_i, \tag{31}$$

where $q_i \equiv p_{Yi}Y_i/n$. Next, from (4)–(7), (30), and (17′), we obtain

$$(\varepsilon-1)(\bar{g}+\rho)\left(\sum_i \frac{\bar{\sigma}_i b_i}{S_i^{\alpha}}\right) - \beta\sum_i \bar{q}_i = 0. \tag{32}$$

Naturally, we also require

$$\sum_i \bar{\sigma}_i = 1. \tag{33}$$

Finally, (15′) implies

$$\bar{q}_i = \frac{\bar{s}_{i1}\bar{e}_1}{T_i} + \bar{s}_{i2}\bar{e}_2. \tag{34}$$

It is straightforward, now, to verify that (31)–(34) imply (25) and (26) when $T_i = S_i = 1$ for $i = 1,2$ (with $\bar{e} = \Sigma_i \bar{e}_i$). This provides a

[11] This statement can be verified using equations (3), (4), and (17′).

consistency check on the extended model with policy instruments.

We consider trade policies first. From (34), the ratio $\bar{q}_1/\bar{q}_2$ satisfies

$$(35)\qquad \frac{\bar{q}_1}{\bar{q}_2}=\frac{\bar{s}_{11}\bar{e}_1+\bar{s}_{12}\bar{e}_2}{\bar{s}_{21}\bar{e}_1/T_2+\bar{s}_{22}\bar{e}_2}.$$

Now, for given expenditure levels $\bar{e}_i$, equations (31)–(33) and (35)—which constitute a system of five equations—provide a solution for $(\bar{g},\bar{\sigma}_1,\bar{\sigma}_2,\bar{q}_1,\bar{q}_2)$. In this system, the trade policy parameters appear only in (35). Therefore, the long-run effects of trade policy depend only on their effects on $\bar{q}_1/\bar{q}_2$, taking into account the induced adjustment in the spending levels $\bar{e}_1$ and $\bar{e}_2$. Moreover, for small trade policies (i.e, with an initial value of $T_2=1$), the spending shares are equal across countries ($\bar{s}_{i1}=\bar{s}_{i2}$), so the effect on $\bar{q}_1/\bar{q}_2$ of changes in the cross-country composition of aggregate spending "washes out."

Further inspection of (35) reveals that an increase in T_2 starting from free trade with $T_2=1$ (i.e., a small import tariff in country 1) unambiguously raises $\bar{q}_1/\bar{q}_2$.[12] A tariff shifts demand by residents of country 1 toward home consumer products, and since relative producer prices do not change in the long run, steady-state relative quantities must adjust. The effect of this change on the steady state is qualitatively the same as for an exogenous increase in world preference for final good 1, such as we studied in the previous section when we varied s_1. Similarly, a small export subsidy in country 2 (a reduction in T_2 to a value slightly below one) biases country 1 demand in favor of foreign final output. So we may apply directly our results from Proposition 1 to state the following:

PROPOSITION 4: *A small import tariff or export subsidy on final goods reduces a country's steady-state share in middle products and R&D. It increases the rate of long-run growth in the world economy if and only if the policy-active country has comparative disadvantage in R&D.*

Commercial policies *do* affect long-run growth rates. They do so by shifting resources in the policy-active country out of the growth-generating activity (R&D) and into production in the favored sector. At the same time, a resource shift of the opposite kind takes place abroad in the dynamic general equilibrium. The net effect on world growth hinges on the identity of the country that favors its consumer-good industry. If import protection or export promotion is undertaken by the country that is relatively less efficient in conducting R&D, then growth accelerates; otherwise, growth decelerates.

Next, we investigate the effects of small subsidies to R&D, introduced from an initial position of *laissez faire*. For these policy experiments, $T_2=1$ before and after the policy change, so the expenditure levels $\bar{e}_i$ cancel from (35). Suppose, first, that both countries apply subsidies at equal *ad valorem* rates; that is, $S_1=S_2=S$. In this case, relative prices of final output do not change across steady states. Therefore, the spending shares $\bar{s}_{ij}$ do not change. In Appendix C of our 1989a working paper, we totally differentiate (31)–(33) and (35) with respect to S to prove the following:

PROPOSITION 5: *A small R&D subsidy by both countries at a common rate increases the rate of long-run growth in the world economy.*

This proposition is not surprising, and corresponds to a similar result for the closed

[12] The easiest way to see this is to write the right-hand-side of (35) as

$$(\bar{p}_{Y1},\bar{p}_{Y2})[\phi_1(\bar{p}_{Y1},T_2\bar{p}_{Y2})\bar{e}_1+\phi_1(\bar{p}_{Y1},\bar{p}_{Y2})\bar{e}_2]/$$
$$[\phi_2(\bar{p}_{Y1},T_2\bar{p}_{Y2})\bar{e}_1+\phi_2(\bar{p}_{Y1},\bar{p}_{Y2})\bar{e}_2],$$

where $\phi_i(\cdot)$ is minus the partial derivative of $v(\cdot)$ from (1) with respect to its ith argument divided by $v(\cdot)$. Then an increase in T_2 with $\bar{p}_{Y1}/\bar{p}_{Y2}$ constant clearly raises demand for final good 1 in country 1 (the first component of the bracketed term in the numerator increases) and lowers the demand there for final good 2 (the first component of the bracketed term in the denominator falls).

economy derived by Romer (1990). Since R&D represents the only source of gains in per capita income in our model, stimulation of this activity promotes growth.

What is more interesting, perhaps, is the effect of a small R&D subsidy in a single country. As for bilateral subsidies, a unilateral subsidy promotes growth by bringing more resources into product development in the policy-active country. But now, relative final-good prices change, so the spending shares in (35) must be allowed to vary unless the utility function has a Cobb-Douglas form. Depending on whether the elasticity of substitution between final products exceeds or falls short of one, this induced change in the pattern of spending can be conducive to or detrimental to growth. Moreover, an R&D subsidy in a single country will alter the relative shares of the two countries in product development. If the subsidy is introduced by the country that is relatively less efficient at performing R&D, this effect too can impede growth. In Appendix D of our 1989a working paper we show, by means of a numerical example using a Cobb-Douglas utility function, that an R&D subsidy introduced by the country with comparative disadvantage in R&D might (but need not) reduce the world's growth rate. We also prove in Appendix C that, for the case of constant spending shares, an R&D subsidy must encourage growth if it is undertaken by the country with comparative advantage in R&D. Thus we have the following:

PROPOSITION 6: *The provision of a subsidy to R&D in one country increases long-run growth if spending shares on the two final goods are constant and the policy is undertaken by the country with comparative advantage in R&D. Otherwise, the long-run growth rate may rise or fall.*

Our results here on the long-run growth effects of government policy have no immediate normative implications, both because we perform only steady-state comparisons and because the initial *level* of utility may differ along alternative growth paths. But in our 1989b paper we conduct a complete welfare analysis for a small country with endogenous growth generated by technological progress as here. We find that the market determined rate of growth is suboptimally low due to the presence of the nonappropriable spillovers in the process of knowledge generation. An R&D subsidy that speeds growth improves welfare for the small country until some optimal growth rate is achieved. Further increases in the subsidy rate reduce welfare even as they accelerate growth. We show also that trade policy need not raise welfare, even if it successfully speeds growth. The explanation for this lies in the presence of a second distortion in our economy, one that stems from the pricing of middle products above marginal cost. The mark-up pricing practiced by intermediate producers gives rise to suboptimal use of middle products in the production of final goods. Trade policy that accelerates growth but reduces the output of middle products in the general equilibrium can be detrimental to welfare. Conversely, commercial policy that augments the output of middle products can improve welfare even if the growth implications of such policy are adverse.

These results do not carry over immediately to the current environment, though the considerations we found to be relevant in our 1989b paper are certainly applicable here as well. Normative analysis in the present two-country setting is complicated by the terms-of-trade effects that arise when policy alters the relative price of the final goods, changes the price of intermediates (if sectoral trade in middle products is not balanced), or varies the interest rate. Because of these various terms-of-trade effects we have been unable thus far to find simple conditions under which growth-enhancing policy by one government raises that country's welfare.

V. Lags in the Diffusion of Knowledge

We have assumed all along that research and development creates as a by-product an addition to the stock of knowledge that facilitates subsequent R&D. Moreover, we supposed that the knowledge so created becomes available immediately to scientists and engineers worldwide. We now relax the latter assumption, in recognition of the fact that privately created knowledge, even if nonappropriable, may enter the public domain via an uneven and time-consuming process. Also, since legal and cultural barriers may inhibit the free movement of people and ideas across national borders, we shall allow here for the possibility that information generated in one country disseminates more rapidly to researchers in the same country than it does to researchers in the trade partner country. After extending the model we will reconsider the effects of trade policies on the steady-state rate of growth.

In place of our earlier assumption that world knowledge accumulates exactly at the rate of product innovation (equation (9)), we suppose now that R&D expenditures contribute to country-specific stocks of knowledge according to

$$(9^\dagger)\quad K_i(t)=\lambda_h\int_{-\infty}^{t} e^{\lambda_h(\tau-t)}n_i(\tau) + \lambda_f\int_{-\infty}^{t} e^{\lambda_f(\tau-t)}n_j(\tau)\,d\tau,$$

where $K_i(t)$ is the stock of knowledge capital at time t in country i. With this specification, the contribution of a particular R&D project to general knowledge is spread over time. At the moment after completion of the project, none of its findings have percolated through the scientific and professional community. After an infinite amount of time has passed, the R&D project makes, as before, a unit contribution to knowledge. After finite time, the contribution lies between these extremes of zero and one, as given by the exponential lag structure in (9†). The parameters λ_h and λ_f (with $\lambda_h \geq \lambda_f$) distinguish within-country and cross-country rates of diffusion.

The introduction of lags in the diffusion of knowledge alters two of the fundamental equations of the model. First, (8) becomes

$$(8^\dagger)\qquad \dot n_i = L_{ni}K_i/a_{Lni}.$$

Second, we have in place of (10),

$$(10^\dagger)\qquad c_{ni}=w_i a_{Lni}/K_i.$$

In a steady state with $\dot n_1=\dot n_2=g$, we have $n_i(\tau)=n_i(t)e^{g(\tau-t)}$, so that

$$(36)\quad K_i(t)=\frac{\lambda_h}{\lambda_h+\bar g}n_i(t)+\frac{\lambda_f}{\lambda_f+\bar g}n_j(t)$$
$$=\left[\frac{\lambda_h}{\lambda_h+\bar g}\bar\sigma_i+\frac{\lambda_f}{\lambda_f+\bar g}\bar\sigma_j\right]n(t)$$
$$\equiv \mu_i(\bar\sigma_1,\bar\sigma_2,\bar g)n(t).$$

So in the steady state, knowledge in each country is proportional once again to the total number of middle products. But the factor of proportionality has become country-specific and endogenous. This means that the steady-state labor-input coefficient for R&D in country i, a_{Lni}/μ_i, also is endogenous; that is, relative productivity in R&D depends now not only on relative natural abilities in performing this activity, but also on relative cumulative experience in research, as summarized by the σ_i's. This consideration leads us to draw a distinction henceforth between *natural* and *acquired* comparative advantage in R&D.

From (36) we see that when $\lambda_h=\lambda_f\to\infty$ (i.e., when diffusion lags are very short), $\mu_1=\mu_2\to 1$, and the extended model reverts to the earlier formulation. For $\lambda_h=\lambda_f$ finite, $\mu_1=\mu_2$, so that the ratio of the natural-plus-acquired productivity parameters for each country is the same as for the natural productivity parameters alone. In this case, the pattern of comparative advantage cannot be reversed by endogenous learning, and all results from before continue to apply. We concentrate here on cases in which the rates of diffusion are *unequal* but the difference between them is small.[13]

[13] A large difference between the within-country and across-country rates of diffusion may imply that, in the steady-state equilibrium, all R&D is carried out by one country. Such specialization, which is common in mod-

We derive the long-run effects of trade policy in the extended model using equations (31)–(33) and (35), but with $S_1 = S_2 = 1$ (no R&D subsidies), with b_i replaced by b_i/μ_i^α (natural plus acquired comparative advantage in place of just natural comparative advantage), and with h_i replaced by $h_i\mu_i$ (natural plus acquired effective labor in place of natural effective labor). For clarity of exposition, we shall also assume for the remainder of this section that the spending shares s_i are constant. This assumption is valid when $u(\cdot)$ takes a Cobb-Douglas form.

The new elements that diffusion lags introduce to the analysis of policy stem from the effects of *relative size* and *demand-side bias*. Before considering these new aspects, let us suppose that labor forces are equal and demand for the two final goods is symmetric. By totally differentiating the system of steady-state equations (see Appendix E of our 1989a working paper), we establish the following:

PROPOSITION 7: *Suppose $L_1 = L_2$, $s_1 = s_2$, $a_{LX1} = a_{LX2}$, and $\lambda_h - \lambda_f > 0$, but small. Then a tariff on imports of final goods in country i raises the long-run growth rate if and only if $a_{Lni} > a_{Lnj}$.*

In this case, the effects of acquired comparative advantage necessarily *reinforce* those of natural comparative advantage. The country that is relatively more productive in creating new blueprints will attain, in the steady-state equilibrium prior to the introduction of policy, a majority share of the world's middle products. By its greater concentration in R&D, it will gain more experience in research and attain a higher steady-state stock of knowledge. Thus, the effects of learning will augment its initial comparative advantage in R&D. When policy is introduced in one country or the other, the implications of the dynamic resource reallocation for the global efficiency of R&D will be all the more significant.

Now suppose that the two countries differ initially only in (effective) size, as measured by h_i. Recall that with equal rates of diffusion, a small tariff in either country does not affect the long-run rate of growth. Now, however, we find the following:

PROPOSITION 8: *Suppose $b_1 = b_2$, $s_1 = s_2$ and $\lambda_h - \lambda_f > 0$, but small. Then a small tariff on imports of final goods raises the long-run growth rate if and only if the policy is introduced by the country with the relatively smaller effective labor force.*

Here, the larger country will come to acquire comparative advantage in R&D, though it starts with none. The reason is as follows. With differential rates of diffusion, knowledge takes on the characteristics of a *local public good*. The larger country will have more (effective) scientists to benefit from this nonexcludable good as its share in world R&D exceeds one half. So it acquires over time a relatively larger knowledge base and hence a relatively more productive corps of researchers. Trade policy that serves to divert resources away from the R&D sector in the larger country once comparative advantage has been established must be detrimental to growth.

The effects of demand-size bias are similar. The country whose good is in relatively greater demand must devote relatively more of its resources to final-goods production. Thus, its R&D sector initially will be smaller. This country develops over time a comparative disadvantage in R&D, as its learning lags that in its trade partner country. Protection in this country will improve world efficiency of R&D and thereby speed growth.

Once we allow for lags in the diffusion of scientific knowledge and differential speeds of diffusion within versus between countries, we find a richer set of possibilities for the long-run effects of trade policy. Comparative advantage continues to play a critical role in determining whether policy in one country will speed or decelerate growth.

els with a national component to increasing returns to scale, necessarily occurs here if static preferences are Cobb-Douglas and $\lambda_f = 0$ (i.e., all spillovers are internal). Then the equations that we have developed to describe the steady-state equilibrium (which presume non-specialization in each country) would not be valid.

But comparative advantage now must be interpreted with care, because it reflects not only natural ability but also the (endogenous) benefits from cumulative experience.[14] Since steady-state productivity in R&D varies positively with the size of the R&D sector, all determinants of the equilibrium allocation of resources to this sector come to be important in the analysis of policy.

VI. Conclusions

In this paper, we have analyzed a dynamic, two-country model of trade and growth in which long-run productivity gains stem from the profit-maximizing behavior of entrepreneurs. We have studied the determinants of R&D, where research bears fruit in the form of designs for new intermediate products and in making further research less costly. New intermediate products permit greater specialization in the process of manufacturing consumer goods, thereby enhancing productivity in final production. In order to highlight the role of endogenous technological improvements as a source of growth we have abstracted entirely from factor accumulation. But Romer (1990) has shown that capital accumulation can be introduced into a model such as the one we have studied without affecting the analysis in any significant way.

The interesting features of our analysis arise because of the assumed presence of *cross-country differences* in efficiency at R&D and manufacturing. Considerations of comparative advantage in research versus manufacturing of intermediate goods bear importantly on the implications of economic structure and economic policy for long-run patterns of specialization and long-run rates of growth. We find, for example, that growth in world resources or improvements in R&D efficiency need not speed the rate of steady-state growth, if those changes occur predominantly in the country with comparative disadvantage in R&D.

Concerning policy, we find for the first time a link between trade intervention and long-run growth. Any (small) trade policy that switches spending toward the consumer good produced by the country with comparative advantage in R&D will cause long-run growth rates to decline. Subsidies to R&D will accelerate growth when applied at equal rates in both countries, but need not do so if introduced only in the country with comparative disadvantage in R&D. When knowledge spillovers occur with a time lag and diffusion is faster within the country of origin than across national borders, comparative advantage becomes endogenous. Once we recognize that comparative advantage can be *acquired* as well as natural, we find a role for country size and demand-size bias in determining the long-run effects of policy.

Our emphasis on comparative advantage in research and development highlights only one channel through which trade structure and commercial policy might affect long-run growth. In other contexts, the trade environment might influence the rate of accumulation of human capital or the rate at which a technologically lagging (less developed) country adopts for local use the existing off-the-shelf techniques of production. Investigation of the links between trade regime and these other sources of growth seems to us a worthy topic for future research.

[14] Endogenous comparative advantage also plays a central role in Krugman's (1987) analysis of commodity-specific learning-by-doing. There, as here, productivity increases with cumulative experience. But each good is produced in only one country in Krugman's model, so long-run comparative advantage is fully determined by the initial pattern of specialization.

REFERENCES

Becker, Gary S. and Murphy, Kevin M., "Economic Growth, Human Capital and Population Growth," *Journal of Political Economy*, forthcoming, 1990.

Corden, W. Max, "The Effects of Trade on the Rate of Growth," in J. Bhagwati et al., eds., *Trade, Balance of Payments, and Growth: Papers in Honour of Charles P. Kindleberger*, Amsterdam: North-Holland, 1971.

Ethier, Wilfred J., "National and International Returns to Scale in the Modern Theory of International Trade," *American Economic Review*, June 1982, *72*, 389–405.

Findlay, Ronald, "Growth and Development in Trade Models," in R. Jones and P. Kenen, eds., *Handbook of International Economics*, Amsterdam: North-Holland, 1984.

Griliches, Zvi, "Issues in Assessing the Contribution of Research and Development in Productivity Growth," *Bell Journal of Economics*, Summer 1979, *10*, 92–116.

Grossman, Gene M., "Explaining Japan's Innovation and Trade: A Model of Quality Competition and Dynamic Comparative Advantage," National Bureau of Economic Research Working Paper No. 3194, December 1989.

______ **and Helpman, Elhanan,** (1989a) "Comparative Advantage and Long-Run Growth," National Bureau of Economic Research Working Paper No. 2809, January 1989.

______ **and** ______, (1989b) "Growth and Welfare in a Small, Open Economy," Woodrow Wilson School Discussion Paper in Economics No. 145, June 1989.

______ **and** ______, (1989c) "Product Development and International Trade," *Journal of Political Economy*, December 1989, *97*, 1261–83.

______ **and** ______, (1989d) "Quality Ladders in the Theory of Growth," National Bureau of Economic Research Working Paper No. 3099, September 1989. (*Review of Economic Studies*, forthcoming.)

______, ______ **and Razin, A.,** eds., *International Trade and Trade Policy*, Cambridge, MA: MIT Press, forthcoming.

Judd, Kenneth, "On the Performance of Patents," *Econometrica*, May 1985, *53*, 567–85.

Krugman, Paul R., "The Narrow Moving Band, the Dutch Disease, and the Competitive Consequences of Mrs. Thatcher: Notes on Trade in the Presence of Dynamic Scale Economies," *Journal of Development Economics*, October 1987, *27*, 41–55.

______, "Endogenous Innovation, International Trade and Growth," *Journal of Political Economy*, forthcoming, 1990.

Lucas, Robert E. Jr., "On the Mechanics of Economic Development," *Journal of Monetary Economics*, July 1988, *22*, 3–42.

Maddison, Angus, "Growth and Slowdown in Advanced Capitalist Economies: Techniques of Quantitative Assessment," *Journal of Economic Literature*, June 1987, *25*, 649–98.

Pasinetti, Luigi, *Structural Change and Economic Growth*, Cambridge: Cambridge University Press, 1981.

Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*, 1002–37.

______, "Endogenous Technological Change," *Journal of Political Economy*, forthcoming, 1990.

Solow, Robert M., "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, August 1957, *39*, 312–20.

The Fundamental Determinants of the Terms of Trade Reconsidered: Long-Run and Long-Period Equilibrium

By William Darity, Jr.*

Ronald Findlay's analysis of the long-run equilibrium (uniform international growth rates) determinants of the terms of trade, in the context of North-South models of trade and growth, is reconsidered when the North is a Keynesian and a Kaleckian economy. Also examined are the determinants of the terms of trade in long-period equilibrium (uniform profit rates), so that the consequences of capital mobility are accounted for. Two surprising results are noted: first, the long-run terms of trade when the North is a Kaleckian economy will be independent of the North's markup and, second, in the long-period it is theoretically possible for the North to raise its markup but experience a deterioration in its own terms of trade. (JEL *110, 411, 441)*

In two important papers, Ronald Findlay (1980, 1981) undertook an extensive examination of the "fundamental determinants" of the terms of trade —fundamental, in the sense, that these are real factors, in dynamic equilibrium, governing the relative price at which commodities exchange in interregional trade. While Findlay (1980, pp. 293–95; 1981, pp. 442–44) also explored short-run factors influencing the terms of trade when they function to clear trade balances, the more intriguing aspects of his papers concern the factors dictating the terms of trade from the long view. Findlay sought to go beyond identification of the short-run direction of change of the terms of trade, looking instead for the central value toward which they must ultimately gravitate.

To identify the long-run value of the terms of trade, Findlay followed the structuralist tradition of investigating their determinants in the context of a North-South trade model. Here the North is the more affluent, industrial region, while the South is the poorer, predominantly agrarian region. The North exports manufactures and the South exports primary products, establishing an asymmetry between the regions via the international pattern of specialization.

One can presume that the pattern arises from old-fashioned comparative advantage. However, the dynamic effects of such a pattern of specialization need not be salutary for Southern growth rates or for the Southern terms of trade. Prebisch (1959) and Singer (1950) both theorized that the South typically would experience slower growth and deteriorating terms of trade, given such an international pattern of specialization. Moreover, for Prebisch and Singer, declining terms of trade served as a virtual index of divergent growth between South and North. While they argued for a secular decline in primary product exporters' terms of trade, the theoretical backdrop to their proposition was essentially short run in character, rooted in the exercise of monopoly power by Northern producers as the terms of trade adjusted to bring about trade balance. They did not advance a theory that was long run, in an analytical sense, to correspond to their proposition about the secular movement in the terms of trade.[1]

*Professor of Economics, University of North Carolina at Chapel Hill. I am grateful to André Burgstaller, Patrick Conway, Amitava Dutt, Bobbie L. Horn, Marc Jarsulic, and especially several anonymous referees for exceptionally useful comments and criticisms.

[1] Of course, there has been a vast amount of controversy about the *empirical* validity of the Prebisch-Singer hypothesis. Findlay (1981) conveniently has summa-

Findlay's endeavor to characterize the long-run terms of trade suggests that they have an eventual "resting point." The short-run or secular trends might have an ultimate position of equilibrium, an equilibrium whose characteristics can inform us about the "fundamental determinants" of the terms of trade. Short-run factors that have the terms of trade equilibrating the trade balance are more ephemeral, in part because, from a Classical perspective, the interaction of supply and demand only offers a value consistent with market clearing (a "temporary" equilibrium) but not a long-run or long-period position (or dynamic equilibrium).

In what follows, fundamental or dynamic equilibrium analyses of the terms of trade are examined in two parts: first, an analysis and extension of the Findlay-styled explanation of the long-*run* terms of trade and, second, an alternative long-*period* explanation of the terms of trade. The distinction between the long run (or uniform growth rate condition) and the long period (or uniform profitability condition) will be explained in greater detail below.

I. The Terms of Trade in the Long Run

The Findlay (1980, pp. 291–93; 1981, pp. 439–54) explanation for the long-run terms of trade is derived from his version of a North-South trade model. Findlay postulates that the North, producing a manufacturing export that serves as an investment good for both regions, is growing neoclassically in smooth Solow fashion. The South, in contrast, is a Lewis unlimited labor economy producing a primary product. Aggregate production functions are characteristically neoclassical in both regions, with well-defined marginal products in both labor and capital.

Findlay also makes Kalecki's extreme (1971) assumption in the South: all savings are from profits; all wages are consumed. The propensity to save out of profits in the South is positive (and large); the propensity to save out of wages in the South is zero. The Southern profit rate, ρ_s, is defined as

$$(1) \qquad \rho_s = \theta\pi'(k_s)$$

$$\theta \equiv \mathbf{P}_s/\mathbf{P}_n,$$

where θ is the terms of trade, the rate at which manufactures exchange per unit of primary products; $\pi'(k_s)$ is the marginal product of capital where the function π is the intensive form of the South's aggregate production function; and k_s is the capital-employment ratio in the South.

If σ is the savings rate out of profits in the South, under the extreme Kaleckian assumption, the South's growth rate, g_s, will be

$$(2) \qquad g_s = \sigma\theta\pi'(k_s).$$

The condition of unlimited labor in the South fixes the real wage in terms of primary products at $\bar{w}$. Profit maximization by Southern employers yields the capital-output ratio, k_s^*, in the South from the following equation that sets the South's real wage equal to its marginal physical product of labor:

$$(3) \qquad \pi(k_s^*) - k_s\pi'(k_s^*) = \bar{w}.$$

The Northern growth rate, g_N, given Solow steady-state growth must be at the exogenous natural rate, n_N:

$$(4) \qquad g_N = n_N.$$

Now Findlay (1980, p. 293; and 1981, pp. 441–42) identifies the *long run* as a condition where the two region's growth rates become the same, or, for "steady-state equilibrium in the world economy," $g_s = g_N$. This constrains the South to save (and accumu-

rized the panoply of objections raised on empirical grounds in one place. Sarkar (1986) provides a recent attempt to defend the Prebisch-Singer hypothesis on the basis of the evidence. Concern here is not with the empirical case for or against Prebisch and Singer's proposition but is devoted exclusively to the question of whether or not there is potential theoretical support for their hypothesis from the perspective of North-South trade that takes the long view.

late) at the North's natural rate of growth. Uniformity of growth rates yields Findlay's expression for the long-run terms of trade:

$$\theta^* = \frac{n_N}{\sigma\pi'(k_s^*)}. \tag{5}$$

This is a very compact and interesting result. Findlay's equation for the terms of trade under balanced growth worldwide has it that the terms of trade are governed by the natural rate of growth in the North, the savings rate out of profit income in the South, technical conditions in the South's export sector, and, implicitly, the real wage in the South. As Findlay (1981, p. 442) observes, the provocative feature of this result is that the terms of trade "[are] completely independent of [the] production function and [the] propensity to save of the North and of the demand for imports in the two regions." The terms of trade function as the relative price that adjusts to bring about balanced growth internationally.

Findlay's result can be considered from a different angle that will prove useful in the ensuing discussion. For macroeconomic equilibrium in the world economy, the following condition must be met:

$$(g_N^I - g_N^S)u + (g_s^I - g_s^S)(1-u) = 0 \tag{6}$$

$$u \equiv K_N/K.$$

Condition (6) says that a weighted sum of the difference between each region's investment and savings rates must equal zero. Weights are given by each region's share in the available capital supplied by the North. Condition (6) is equivalent to the statement that, in a global sense, investment must equal savings.[2]

There are two circumstances under which (6) will hold. First, the investment and savings rates could equalize separately in the North and the South. Second, the two regions could run offsetting "surpluses" and "deficits," with the surplus region transferring capital to the region experiencing the deficit. In the first case, it is reasonable to treat the capital or manufactured good as internationally immobile after installation in either region; capital is treated as fixed. In the second case, the effects of capital mobility must be taken into account; capital can be treated as circulating.

From the standpoint of condition (6), Findlay adopts the first type of circumstance, assuming for the North that $g_N^I = g_N^S$ and for the South that $g_s^I = g_s^S$. The North's saving rate, given its character as a Solow economy, is given by $sa(k_N)$, where s is the propensity to save, the function $a(k_N)$ is the average product of capital, and k_N is the capital intensity in North's production. The North's investment rate, implicitly, is given by its natural rate of growth, n_N. The South, sharing the North's growth rate in the long run, must also invest at rate n_N, while the South's saving rate, as noted above, is dictated by $\sigma\theta\pi'(k_s^*)$. It is equalization of investment and savings rates in the South that provides a basis for derivation of Findlay's result for the long-run terms of trade, $\theta^* = n_N/\sigma\pi'(k_s^*)$.

Findlay's assumption of full employment growth in the North eliminates Northern savings behavior and technology from being determinants of the long-run terms of trade. However, with unemployed labor in both regions, matters can change sharply, although demand elasticities will remain irrelevant. Note also in what follows, similar to Findlay, Laursen-Meltzer (1950) effects are ignored—so that savings rates in both regions do not depend upon the terms of

[2] World macroeconomic equilibrium requires that the international sum of investment and sum of savings must cancel, or $(I_N - S_N)+(I_s - S_s)=0$, where $I_i(i = N, S)$ is investment in each region and $S_i(i = N, S)$ is savings in each region. The global macroeconomic equilibrium condition can be transformed into growth rates by multiplying each bracketed term as follows: $(I_N - S_N)K_N/K_N + (I_s - S_s)K_s/K_s = 0$. Since $I_i/K_i = g_i^I$ and $S_i/K_i = g_i^S$, the preceding expression can be written as $g_N^I - g_N^S)K_N + (g_s^I - g_s^S)K_s = 0$. Defining $\hat{u} = K_N/K$, the North's share in the global capital stock, means that we can obtain the expression in equation (6): $(g_N^I - g_N^S)u + (g_s^I - g_s^S)(1-u) = 0$. This equilibrium condition undergirds much of Taylor's (1983) work on North-South trade and growth.

trade. In fact, savings rates are assumed fixed and immutable in both North and South for posterity.

After Sen (1963) and Taylor (1981; 1983, p. 181), the North can be given a more Keynesian cast by dropping the assumption of full employment/natural rate growth and introducing, instead, an independent investment function. Northern investment can be treated as an increasing function of the North's rate of profit, ρ_n:

$$(7) \qquad g_N^I = i_0 + i_1\rho_n = i_0 + i_1 f'(k_N).$$

The parameters i_0 and i_1 are positive constants, and $f'(k_N)$ is the marginal product of capital in the North. Setting investment equal to savings in the North yields the equilibrium value for the North's capital intensity, k_N^{**}:

$$(8) \qquad g_N^I = g_N^S = \rightarrow i_0 + i_1 f'(k_N^{**}) = sa(k_N^{**}).$$

In Figure 1 the determination of k_N^{**} is displayed visually. To obtain an economically meaningful solution for k_N^{**}, the savings schedule must intersect the investment schedule from above.[3]

Long-run equilibrium necessitates that the South grow at the same pace as the North. Therefore, either $i_0 + i_1 f'(k_N^{**})$ or $sa(k_N^{**})$ has to be set equal to the South's investment rate, and then, in turn, the South's saving rate must adapt to the North's growth rate via adjustment in the terms of trade. Using $sa(k_N^{**})$ for the North's growth rate, the long-run terms of trade now will become

$$(9) \qquad \theta^* = \frac{sa(k_N^{**})}{\sigma\pi'(k_s^*)}.$$

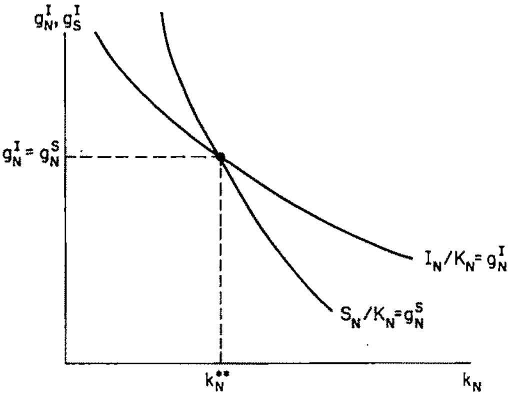


FIGURE 1

Note that the equilibrium value of the North's capital intensity, k_N^{**} will not, in general, be the same as the capital intensity that prevailed under Solow full-employment growth, k_N^*, and $sa(k_N^{**})$ will not, in general be the same as the natural growth rate, n_N. Both savings propensities now matter —σ out of profits in the South and s out of total income in the North—in establishing the long-run value of θ. Technical conditions in both regions also matter now. Again implicit in the determination of k_s^* is the Lewis South's fixed real product wage. Correspondingly, determination of k_N^{**} in the Keynesian North means that implicitly the long-run terms of trade also depend upon the parameters of the investment function.

Keep in mind that only profits are saved in the South, while some of both categories of income are saved in the North. Even this latter asymmetry need not be maintained if we generalize the North's savings function in Kaldor (1955) fashion. Suppose there are different propensities to save out of profits and wages in the North, s_p out of profits and s_w out of wages. For the time being retain Findlay's assumptions that the rate of profit can be identified with the marginal product of capital. The North's savings rate will be

$$(10) \qquad g_N^S = (s_p - s_w) f'(k_N) + s_W a(k_N).$$

If we maintain the assumptions that savings and investment separately must equalize in each region, that North and South grow at the same rate, and that the Keynesian North has an independent investment function of

[3] If, for example, the intensive form of the North's production function were Cobb-Douglas or $f(k_N) = k_N^a$, the equilibrium value of the capital labor ratio would be $k_n^{**} = [i_0/(s - i_1 a)]^{[1/(a-1)]}$. To ensure that k_n^{**} is a positive number, $s > i_1 a$, or the savings schedule must be steeper than the investment schedule.

the form $i_0 + i_1 f'(k_N)$, the long-run terms of trade will be

$$(11)\quad \theta^* = \frac{(s_p - s_w)f'(k_N^{**}) + s_w a(k_N^{**})}{\sigma\pi'(k_s^*)}.$$

In the special case where $s_w = 0$, again the extreme case of a Kalecki (1971) savings function, (11) simplifies to

$$(12)\qquad \theta^* = \frac{s_p f'(k_N^{**})}{\sigma\pi'(k_s^*)}.$$

Now the factors influencing the terms of trade in both regions appear to be symmetrical—the invariant savings rates out of profits and production functions or, some might say, tastes (for foregone consumption) and technology. The remaining asymmetry involves the special role of the North's investment function, which must become, *de facto*, the South's investment function because steady-state growth in the world economy means the South's growth rate conforms to the North's.[4]

II. Markup Pricing in the North and the Long-Run Terms of Trade

Thus far, the results examined have been derived under the assumption of neoclassical production functions and marginal product factor pricing in both regions. And, thus far, the results derived are consistent with any economically meaningful distribution of international capital between North and South; that is, u can take any value between zero and one. How might the long-run terms of trade be characterized under alternative assumptions about technology and distribution theory?

To push back toward the types of asymmetries that enliven North-South models, treat the North as following the dictates of a markup pricing rule in Kaldor-Kalecki-Weintraub fashion (see Reynolds, 1987, for a useful exposition). This also corresponds to Prebisch (1959) and Singer's (1950) idea that the manufactured goods producer exercises monopoly power in pricing. It also is similar to Taylor (1981), who confined markup pricing to the North.

With a markup, m_N, over unit costs, the North's manufactured goods' price will be

$$(13)\quad P_N = (1 + m_N)\left[(w_N L_N / K) + (P_N K_N / K)\right].$$

The markup is assumed to be stable for institutional reasons discussed in Reynolds (1987, pp. 37–62).[5]

In (13) P_N is the price of the North's good, the capital good; w_N is the nominal wage; L_N is the level of employment; Y_N is the real quantity of the North's output; and K_N is the North's capital stock.

If the North's profit rate is defined as the ratio of the surplus above expenditures to the value of the capital stock, using (13) markup pricing implies that the Northern rate of profit will be

$$(14)\quad \rho_N = \frac{m_N(w_N L_N + P_N K_N)}{P_N K_N} = \left(\frac{m_N}{1 + m_N}\right)\left(\frac{1}{u}\right) = M_N / u$$

$$u \equiv K_N / K \qquad M_N \equiv \frac{m_N}{1 + m_N}.$$

The term M_N in the numerator on the far right-hand side of (14) is the ratio of the markup over one plus the markup; the term

[4] There is still a further difference that lurks just below the surface. Northern output is a basic good, while Southern output is nonbasic in a Sraffa (1950) sense. I am grateful to André Burgstaller for pointing this out to me.

[5] The markup pricing mechanism often is introduced in analyses where firms are presumed to have excess capacity. Here, however, it is assumed that although there is generally less than full employment of labor, the capital stock is utilized fully. Peter Reynolds (1987, p. 54) suggests that whereas markup pricing could give way to supply and demand determination of prices once full capacity is reached, "this is the exception rather than the rule" in Kaleckian analysis.

u in the denominator is, again, the North's share in the total available capital.

Now, if North's investment increases with profitability, the investment function will take the following form:

$$(15)\quad g_N^I = i_0 + i_1\rho_N = i_0 + i_1\cdot(M_N/u).$$

Assume, further, that savings in the North are exclusively drawn from profits so that

$$(16)\quad g_N^S = s_p\rho_N = s_p\cdot(M_N/u).$$

When investment and savings rates equalize in the North, they no longer determine the North's capital intensity. Instead, they determine the North's share in global capital. Equality between the right-hand sides of (15) and (16) yields the following equilibrium value for the Northern proportion of international capital:

$$(17)\quad u^* = \left(\frac{s_p - i_1}{i_0}\right)\left(\frac{m_N}{1+m_N}\right)$$

$$= \frac{(s_p - i_1)}{i_0}\cdot M_N.$$

Note, again, for an economically meaningful value of u the savings schedule must be steeper than the investment schedule.

Now if the South is constrained to grow at the North's growth rate to achieve long-run equilibrium, then the North's growth rate, $s_p\cdot(M_N/u^*)$, will serve as the South's investment rate. To have $g_s^I = g_s^S$ means that

$$(18)\quad s_p\cdot(M_N/u^*) = \sigma\theta\pi'(k_s^*),$$

where, again, the terms of trade adjust to achieve the equality.

Substitution of (17) into (18) throws up an intriguing result for the long-run terms of trade:

$$(19)\quad \theta^* = \frac{s_p i_0}{(s_p - i_1)[\sigma\pi'(k_s^*)]};$$

Now the terms of trade depend on the parameters of the savings and investment functions in the North as well as the savings propensity, technical conditions, and the real wage in the South. The numerator is the product of the North's propensity to save out of profits and the autonomous component of the North's investment function. What is striking is the independence of the long-run terms of trade from the North's markup. Of course, the size of the markup does affect directly the share of capital located in the North. A higher markup, amounting to a regressive redistribution of income against labor in the North, would mean a higher equilibrium value for u.

Up to this point, condition (6) has been met by assuming that investment matches savings separately for each region and that the South shares the North's growth rate. This need not be the case, however. Indeed, it will not be the case if the South also has an independent investment function of a similar type as the North's.

Investment in the South can be specified as an increasing function of South profitability in parallel fashion with the North's investment function:

$$(20)\quad g_s^I = h_0 + h_1\rho_s = h_0 + h_1\theta\pi'(k_s^*).$$

Since k_s^* is determined by profit maximization and the presence of a fixed real wage in the unlimited labor South, it is only by sheer accident that the parameters of the South's investment and savings schedules will align to equalize g_s^I and g_s^S. Consequently global macroeconomic equilibrium will hold when the following condition obtains:

$$(21)\quad [i_0u + (i_1 - s_p)(M_N)]$$

$$+ [h_0 + (h_1 - \sigma)\theta\pi'(k_s^*)](1-u) = 0,$$

where investment and savings are *not* separately equal in each region.

The difficulty now is that the system is overdetermined. Condition (21) offers a single equation, but there are two unknowns, θ and u. Moreover, condition (21) implies a transfer of capital from one region to the other, so that even after installation, the available capital is mobile. It is natural to assert that the available capital shifts in response to profit rate differentials between the regions. It also is natural to consider the North to be the "surplus" capital or excess savings region, so that the term $[i_0 u + (i_1 - s_p) M_N]$ is negative, while the South is the "deficit" capital region or excess investment region, so that the term $[h_0 + (h_1 - \sigma)\theta\pi'(k_s^*)]$ is positive. This suggests an alternative characterization of dynamic equilibrium developed in the next and final section of the paper.

III. The Terms of Trade in the Long Period

Findlay's long run is one of balanced steady-state growth. There is, however, at least one alternative procedure for characterizing equilibrium growth from a long view. This is the long-period concept from Classical political economy (Ricardo) where equilibrium prevails when rates of profit become the same in all activities.[6] In place of uniformity of growth rates between North and South, or the condition $g_N = g_s$, substitute uniformity of profit rates, or $\rho_N = \rho_s$.

These conditions are not generally compatible. If both regions possess independent investment functions that are not identical and propensities to save that are not identical, uniformity of profit rates will not correspond to uniformity of growth rates.[7] A global economy in long-period equilibrium can display persistent divergence in growth rates between North and South—a dynamic equilibrium with uneven development.

[6] Implicitly, the model of Arthur Lewis's (1969, pp. 17–22) Wicksell lectures — the same as the model of trade between the temperate zone and the tropics near the end of his (1954, pp. 182–84) justifiably famous paper on growth with surplus labor—adopts a uniform profit rate condition to characterize the equilibrium terms of trade. The international profit rate simply equalizes at zero since he assumes the perfectly competitive zero profit condition. His qualitative results would not alter if he assumed positive profit rates as long as, once again, they are uniform across all sectors (see Darity, 1989).

Stable relative productivities between North and South agriculture and North and South export commodities lead to a determinant equilibrium expression for the terms of trade. If productivities are changing but the pattern of specialization is stable, then Prebisch-Singer results can follow. Lewis (1969; also see Findlay, 1981, p. 433) asserts that empirically productivity has risen fastest in developed country agriculture, next fastest in developed country manufacturing, slower in developing country primary product making, with virtually no technical progress in developing country agriculture.

David Evans (1987) says that the empirical case does not unambiguously favor Lewis's claims of a sectoral bias in technical change in favor of temperate zone (Northern) agriculture. Evans also considers demand-side effects, asserting that a key role must be assigned to Engel curve effects in the determination of the terms of trade. However, Bardhan (1982) has sought to demonstrate that the basic Lewis results remain intact even when demand considerations are brought on the scene in a full general equilibrium treatment. As Lewis (1969, p. 22) himself observed, "the terms of trade are tied rigidly by productivities only for so long as it pays both countries to produce both food and one other commodity. As soon as productivities move outside these limits, the terms of trade cease to be tied rigidly."

André Burgstaller (1987) recently has provided a clever speculative history (and forecast?) of how the terms of trade might unfold under various scenarios as the pattern of specialization alters between "England" and "India." In Burgstaller's long run—the fifth and final stage of his Classical model of trade and development—India becomes the industrial center and England an agrarian nation, reversing their initial roles. Of course, the terms of trade continuously will move against the "South," but here the South formerly was the North!

[7] In general in a Findlay long run, rates of profit will not equalize. For example, in Findlay's case when $\theta^* = n_N / \sigma\pi'(k_s^*)$, the rate of profit in the South simplifies to the ratio n_N / σ. This will only be equal to the North's rate of profit, $f'(k_N^*)$ by a conscious and nonarbitrary selection of values for n_N and σ.

Or, consider that since $u = K_N / K$ is the North's share in world capital, $\hat{u} = (1-u)(\hat{K}_N - \hat{K}_s)$. The North's capital share will become stationary (or $\hat{u} = 0$) under two circumstances. Either $u = 1$, implying that the capital good is used exclusively in the North and South production disappears altogether, or $\hat{K}_N = \hat{K}_s$, implying that the Findlay long-run condition of balanced international growth must hold. When each region has independent investment functions dictating their rates of growth, that is, $\hat{K}_N = i_0 + i_1 \rho_N$ and $\hat{K}_s = h_0 + h_1 \rho_s$, growth rates only will coincide when profit rates diverge—unless the investment functions happen to be identical for North and South. Therefore, in this final section an alternative specification for the $\hat{u}$ equation is advanced, consistent with the notion that capital continues to be moved between regions until profit rates equalize.

The shift rule for movements of available capital will be based upon a profitability gap between the regions.[8] The variable u has been defined as the ratio K_N/K, the North's share of world capital. If capital is fully mobile, this capital share will vary in response to the profit rate differential between the regions.

Therefore, the percentage rate of change of u, represented by the careted variable below, responds positively when $\rho_N > \rho_s$ and negatively when $\rho_N < \rho_s$. The international distribution of capital ceases to change when profit rates become uniform, so that u takes on a fixed value. Therefore, the dynamic equation capturing the presence of capital mobility is

$$\hat{u} = q\cdot(\rho_N - \rho_s). \tag{22}$$

The parameter q is a positive constant that can be interpreted as a reaction-response coefficient. The larger its value, the more rapid the convergence between profit rates; in the limit, when $q = \infty$ the convergence would be instantaneous.

First, consider the setting with independent investment functions in both regions under joint marginal product factor pricing. The key is how to determine the capital intensity in the North, k_N, although both the assumption of natural rate growth and the assumption that $g_N^I = g_N^S$, made above, have been dropped. Growth in the North's capital stock is governed by the rate of investment there, g_N^I. The percentage rate of increase of North's employment can be considered a negative function of the real wage in the North. Therefore, the specification of the percentage growth rate of the North's capital intensity is

$$\hat{k}_N = \hat{K}_N - \hat{L}_N = i_o + i_1\rho_N - [x_0 - x_1(w_N/P_N)]. \tag{23}$$

The parameters x_0 and x_1 are positive constants in what can be interpreted as a dynamic labor demand function. The ratio w_N/P_N is the real wage in the North.

Under marginal product factor pricing the rate of profit, ρ_N, will be the marginal product of capital in the North, $f'(k_N)$, and the real wage, w_N/P_N, will be the marginal product of labor in the North. An equilibrium value of the capital intensity, k_N^0, can be derived from the stationary condition for equation (23):

$$\hat{k}_N = i_0 + i_1 f'(k_N^0) - x_0 + x_1[f(k_N^0) - k_N^0 f'(k_N^0)] = 0. \tag{24}$$

Figure 2 displays the equilibrium level of the North's capital intensity at the intersection of the $\hat{K}_N$ and $\hat{L}_N$ schedules.

The equilibrium value of the South's capital intensity still is determined by the presence of the fixed real wage there. With the equilibrium value of the North's capital intensity in hand, it is straightforward to ob-

[8] Burgstaller and Saveedra-Rivano (1984)—also discussed in Findlay (1984, pp. 226–29)—modify the Findlay North-South model and introduce capital mobility. They manage to construct a model where simultaneously long-run and long-period conditions obtain. For their version of a world with a Solow North and a Lewis South, the terms of trade must satisfy both Findlay's long-run condition of balanced international growth and also a long-period condition that the terms of trade must equal the ratio of the North's marginal physical product of capital to the South's marginal physical product of capital. This implies, in turn, that the following equality must hold: $n/\sigma\pi'(k_s) = f'(k_N)/\pi'(k_s)$, which simply means that balanced growth must obtain in the North between its labor force and its capital stock, a condition Solow growth ensures in the first place.

On the other hand, in Burgstaller (1985) the equilibrium terms of trade are determined solely on the basis of the long-*period* criterion. Capital is treated as a wages fund rather than a produced means of production. The mobility of advances to labor across regions in response to profit rate differentials assures eventual equalization of returns. Under long-*period* conditions Burgstaller derives the following result for the terms of trade: $\theta^* = (w_s q_N)/(w_N q_s)$, where each w_i is the wage rate in the corresponding region and each q_i is the output-labor ratio in each region. As w_i goes up or q_i goes down, production costs in each region rise. Burgstaller directly obtains a strictly Classical result; the terms of trade improve as costs of production rise, and they deteriorate as costs of production decline. With capital as a produced means of production the Classical result appears only indirectly; that is, as the markup rises in a particular region, labor's share in income falls, hence *relative* labor costs fall, and the terms of trade decline for that region.

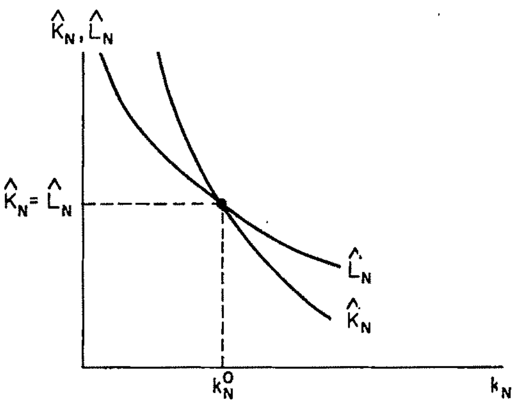


Figure 2

tain the long-period terms of trade by finding the stationary value for equation (22), $\hat{u} = 0$:

$$\theta^* = f'(k_N^0)/\pi'(k_s^*). \tag{25}$$

When condition (25) is attained there is no incentive to transfer capital between the regions. The long-period terms of trade, in this case, are simply the ratio of the respective North and South marginal products of capital.

The macroeconomic equilibrium condition, (21), now can be used to solve for u, yielding the equilibrium international distribution of capital. The long-period value of the share of capital ultimately located in the North depends upon all the parameters of the investment and savings functions in both regions and, by using (25), it depends upon the technical conditions in the North:

$$u^* = \frac{h_0 + (h_1 - \sigma)f'(k_N^0)}{\{[h_0 + (h_1 - \sigma)f'(k_N^0)] - [i_0 + (i_1 - s_p)f'(k_N^0)]\}}. \tag{26}$$

The prior assumption that the North is the capital surplus region and the South is the capital deficit region ensures that u^* is positive in sign and takes a value between zero and unity.

Next consider the case where the North practices markup pricing, exercising some degree of monopoly power in trade, while the South continues to be a marginal product factor pricing region. This is, again, the analogue to the Prebisch-Singer case that emerges from this type of analysis. Under asymmetric international pricing of this sort, when $\hat{u}$ from (22) is set equal to zero, the following relationship between θ and u emerges:

$$u = M_N/\theta\pi'(k_s^*) \tag{27}$$

$$M_N \equiv [m_N/(1+m_N)].$$

Now the macroequilibrium condition, (21), provides the needed second equation in the same two variables. In principle, the combination of equations (27) and (21) offers the route to a long-period solution for the value of θ and u under asymmetric output pricing between North and South. However, because of the structure of (21) it is not possible to derive a closed form solution for either θ or u. The strategy pursued here is to use (27) to substitute for u in (21) and then totally differentiate the expression that remains to investigate the partial derivative $\partial\theta/\partial M_N$. Does a rise in the markup in the North, as Prebisch and Singer suggest, improve the terms of trade for the North? The relevant expression for assessing this question is

$$\frac{\partial\theta^*}{\partial M_N} = \frac{\{(i_0 - h_0)[1/\theta^*\pi'(k_s^*)] + (i_1 - s_p) + (\sigma - h_1)\}}{\{(i_0 - h_0)[M_N/\pi'(k_s^*)] + (\sigma - h_1)[M_N/\pi'(k_s^*)]\}}. \tag{28}$$

Because the North is the capital surplus region, while the South is the capital deficit region, it is reasonable to treat both of the terms $i_1 - s_p$ and $(\sigma - h_1)$ as negative. If the rate of autonomous investment in the North is smaller than the rate of autonomous investment in the South, that is, if $i_0 < h_0$, then a paradoxical result occurs. An increase in the degree of monopoly in the North leads to a *deterioration* in the North's terms of trade.

However, if the rate of autonomous investment in the North exceeds the rate in the South and the product $(i_0 -$

h_0)[$1/\theta\pi'(k_s^*)$] is sufficiently large to make the numerator positive while the denominator remains negative, it is possible to have a circumstance where an increase in the North's markup—corresponding to a regressive redistribution of income—leads to an improvement in the North's terms of trade.[9] This is more likely to be the case if the initial long-period value of θ^* is smaller than unity. The Prebisch-Singer result can emerge in the long period, although it is absent in the long run, when the terms of trade under asymmetric international pricing are independent of the North's markup.

Finally, consider the symmetric case where producers in both regions possess some degree of monopoly power and utilize a markup pricing strategy. In this instance, while the rate of profit for the North is given by (M_N/u), the South's rate of profit will be

$$\rho_s = M_s\alpha/\theta \tag{29}$$

$$M_s = [m_s/(1+m_s)]$$

$$\alpha \equiv K_s/y_s.$$

The term α represents the output-capital ratio in the South. To simplify the exposition impute a fixed coefficients technology to the South, so that α is a constant.

When the North's share of the global capital stock stabilizes, so that $\hat{u}=0$, now it implies the following relationship between u and θ:

$$u = (M_N/M_s)(\theta/\alpha). \tag{30}$$

Since again (21) cannot be used to solve directly for a closed form solution for θ, we pursue a similar strategy to that employed in the asymmetric pricing case. After substituting u from (30) into equation (21), the equation is differentiated totally and the partial derivatives $\partial\theta/\partial M_N$ and $\partial\theta/\partial M_s$ are derived:

$$\frac{\partial\theta^*}{\partial M_N} = \frac{\{(i_0/M_N)(\theta^*/\alpha)+(i_1-s_p) - [h_0+(h_1-\sigma)\rho_s^*][(1/M_s)(\theta^*/\alpha)]\}}{\{-i_0\alpha(M_N/M_s)+(\sigma-h_1)\alpha M_s(1-u^*) + [h_0+(h_1-\sigma)\rho_s^*][(M_N/M_s)/\alpha]\}} \tag{31}$$

$$\frac{\partial\theta^*}{\partial M_s} = \frac{\{-i_0(M_N/M_s^2)(\theta^*/\alpha) + [(h_1-\sigma)(\alpha/\theta)](1-u^*)\}}{\{-i_0\alpha(M_N/M_s)+(\sigma-h_1)\alpha M_s(1-u^*) + [h_0+(h_1-\sigma)\rho_s^*][(M_N/M_s)/\alpha]\}}. \tag{32}$$

Inspection of equations (31) and (32) indicates that both the Prebisch-Singer outcome—a rise in its markup improves a region's terms of trade—and the paradoxical outcome—a rise in its markup causes a region's terms of trade to deteriorate—are theoretically possible under a wide range of parameterizations. This is especially true given the presence of the parameter α, the output-capital ratio in the South.

There are several lingering difficulties. Are centers of gravity for the terms of trade empirically as well as analytically of interest? What sets of trading partners actually fit the various asymmetries assigned to North and South in this discussion? How does Lewis's observation about changing relative productivities and patterns of specialization influence the theoretical outcomes for the terms of trade? How do Marshall's (1920, p. 601) worrisome "complex actions and relations of credit" (and finance) impinge upon the conclusions reached here; is either the long run or long period truly independent of monetary considerations? These questions are not Solow's (1956, pp. 93–94)

[9] The Prebisch-Singer case is the same as the results obtained by Amitava Dutt (1988) and Patrick Conway and myself (1988) in models of the determination of the terms of trade in the *short run*, that is, when the terms of trade serve to bring about balance of trade equilibrium. However, this case is the opposite of what I obtained in Darity (1989), where I found that the paradoxical outcome of a rising markup in the North leading to a deterioration in the terms of trade for the North in both the long run and long period was the *general* result! This earlier unambiguous finding appears to be attributable to the lack of the explicit dynamic structure used in this paper that combines the $\hat{u}=0$ equation with the world macroeconomic equilibrium condition.

"cobwebs" to be brushed lightly aside; these are serious issues to be pursued eagerly in further research that extends the search for a fundamental explanation of the terms of trade between developed and developing countries.

REFERENCES

Bardhan, Pranab K., , "Unequal Exchange in a Lewis Type World," in Mark Gersovitz et al., eds., *The Theory and Experience of Economic Development*, London: Allen & Unwin, 1982.

Burgstaller, André, "Industrialization, Deindustrialization, and North-South Trade," *American Economic Review*, December 1987, *77*, 1017–18.

______, "North-South Trade and Capital Flows in a Ricardian Model of Accumulation," *Journal of International Economics*, May 1985, *18*, 241–60.

______ **and Saveedra-Rivano, N.,** "Capital Mobility and Growth in a North-South Model," *Journal of Development Economics*, Nos. 1, 2, 3, May-June-August 1984, *15*, 213–37.

Conway, Patrick and Darity, William, Jr., "Growth and Trade with Asymmetric Returns to Scale: A Model for Nicholas Kaldor," mimeo., University of North Carolina at Chapel Hill, 1989.

Darity, William, Jr., "The Terms of Trade from the Long View," in Paul Davidson and Jan Kregel, eds., *Macroeconomic Problems and Policies of Income Distribution*, Hampshire: Gower Publishing Group, 1989, 157–73.

Dutt, Amitava K., "Monopoly Power and Uneven Development: Baran Revisited," *Journal of Development Studies*, January 1988, *24:2*, 161–76.

Evans, David, "The Long-Run Determinants of North-South Terms of Trade and Some Recent Empirical Evidence," *World Development*, May 1987, *15*, 657–71.

Findlay, Ronald, "The Terms of Trade and Equilibrium Growth in the World Economy," *American Economic Review*, June 1980, *70*, 291–99.

______, "The Fundamental Determinants of the Terms of Trade," in Sven Grassman and Erik Lundberg, eds., *The World Economic Order: Past and Prospects*, London: Macmillan, 1981, 425–57.

______, "Growth and Development in Trade Models," in R. W. Jones and P. B. Kenen, eds., *Handbook of International Economics*, Vol. 1, Amsterdam: Elsevier Science Publishers, 1984, 185–236.

Kaldor, Nicholas, "Alternative Theories of Distribution," *Review of Economic Studies*, 1955, *23*, 83–100.

Kalecki, Michal, *Selected Essays on the Dynamics of the Capitalist Economy*, Cambridge: Cambridge University Press, 1971.

Laursen, S. and Meltzer, L. A., "Flexible Exchange Rates and the Theory of Employment," *Review of Economics and Statistics*, November 1950, *32*, 281–99.

Lewis, W. A., "Economic Development with Unlimited Supplies of Labour," *The Manchester School*, May 1954, *28*, 139–91.

______, *Aspects of Tropical Trade*, Stockholm: Almqvist Wicksell, 1969.

Marshall, Alfred, *Principles of Economics: An Introductory Volume*, 8th ed., London: Macmillan, 1920.

Prebisch, Raul, "Commercial Policy in the Underdeveloped Countries," *American Economic Review*, May 1959, *49*, 251–73.

Reynolds, P. J., *Political Economy: A Synthesis of Kaleckian and Post Keynesian Economics*, New York: St. Martin's Press, 1987.

Ricardo, David, *Principles of Political Economy and Taxation*, P. Sraffa and M. Dobb, eds., Cambridge: Cambridge University Press, 1951.

Sarkar, Prabirjit, "The Singer-Prebisch Hypothesis: A Statistical Evaluation," *Cambridge Journal of Economics*, December 1986, *10*, 355–71.

Sen, Amartya, "Neo-Classical and Neo-Keynesian Theories of Distribution," *Review of Economic Studies*, 1963, *39*.

Singer, Hans, "The Distribution of Gains Between Investing and Borrowing Countries," *American Economic Review*, May 1950, *40*, 473–85.

Solow, R. M., "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics*, February 1956, *70*, 65–94.

Sraffa, Piero, *Production of Commodities by*

Means of Commodities, Cambridge: Cambridge University Press, 1960.

Taylor, Lance, "South-North Trade and Southern Growth: Bleak Prospects from a Structuralist Point of View," *Journal of International Economics*, November 1981, *11*, 589–601.

______, *Structuralist Macroeconomics: Applicable Models for the Third World*, New York: Basic Books, 1983.

Weintraub, Sidney, "Generalizing Kalecki and Simplifying Macroeconomics," *Journal of Post Keynesian Economics*, Spring 1979, *1*, 101–16.

The Case of the Vanishing Revenues: Auction Quotas with Monopoly

By Kala Krishna*

This paper examines the effects of auctioning quota licenses when monopoly power exists. Here the sales of licenses will never raise any revenue if domestic and foreign markets are segmented. More surprisingly, the inability to raise revenue is shown to persist even when arbitrage across markets is possible as long as the quota is not too far from the free trade import level. This suggests that existing revenue estimates from auctioning quota licenses, which are based on the assumption of competition, are upwardly biased. It also makes it likely that quotas implemented by auctioning licenses, even when set optimally, have adverse welfare effects. (JEL *422)*

In this paper I examine the case for auction quotas when there is a foreign monopolist. A companion paper deals with oligopolistic markets.[1]

One of the most common criticisms of voluntary export restrictions (VERs) and of the way quotas are currently allocated is that they allow foreigners to reap the rents associated with the quantitative constraint. It has been suggested that auctioning import quotas would be a remedy for this.

A Congressional Budget Office (CBO) memorandum estimates quota rents possible in 1987 for a group of industries to be 3.7 billion dollars.[2] It compares this to the C. Fred Bergsten, Ann Kimberly, Elliott Schott, and Wendy Takacs (1987) estimate made for the Institute for International Economics (IIE) of 5.15 billion. More recently, David Tarr (1989), using a general equilibrium model, estimated these rents in textiles, steel, and autos to be even higher —about 14 billion dollars in 1984. Takacs (1987) points out that proposals to auction quotas have become increasingly frequent.[3] She states: "Commissioners Ablondi and Leonard of the U.S. International Trade Commission (ITC) recommended auctioning sugar quota licenses in 1977. The ITC recommended auctioning footwear quotas in 1985. Studies by Gary Hufbauer and Howard F. Rosen (1985) and Robert Z. Lawrence and Robert E. Litan (1985) suggested auctioning quotas and earmarking the funds for trade adjustment assistance."[4]

Despite the importance of the issues involved, the intuition behind such statements and the procedure used in the estimation are based on models of perfect competition. In such models, the level of the quota determines the domestic price, and the difference between the domestic price and the world price determines the price of a license when auctioned. If the country is small, then the world price is given. If the country is large, then the world price does change with a quota. How it changes is

*Associate Professor of Economics, Harvard University. I am grateful to the World Bank and the Center for International Affairs at Harvard for Research Support, to Vijay Krishna for asking the question, to Susan Collins for the title, to Kathleen Hogan and Phillip Swagel for able research assistance, to Elhanan Helpman, David Richardson, Gene Grossman, M. Ben Zvi, Lawrence Summers, David Tarr, Jagdish Bhagwati, and Wendy Takacs for comments. I am particularly grateful to a referee for helping me revise the paper, though the usual disclaimer applies.

[1] See Krishna (1988a).

[2] Memorandum of February 27, 1987, from Stephen Parker on revenue estimates for auctioning existing import quotas (publicly circulated).

[3] The interested reader should consult Bergsten et al. (1987) and Takacs (1987) for a historical and institutional perspective on work in this area.

[4] See Takacs (1987), fn. 7.

determined by supply and demand conditions in the world market.

However, when markets are imperfectly competitive, as they are thought to be in the market for autos, this analysis is misleading.[5] In such environments, prices are chosen by producers so that there is no supply curve, and the response of producers to the constraint must be taken into account when determining the price of a license when it is auctioned off. For example, if the response of profit-maximizing producers is to adjust their prices so that there is no benefit to be derived from owning a license to import, its auction price must be zero!

Therefore, the question that needs to be addressed concerns the behavior of producers in response to quantitative constraints, and the impact of this on the price of a license. There are two policy questions that need to be addressed. First, should existing licenses be auctioned off? Second, if quotas are set at their optimal level, can they be welfare-improving over free trade? There has been relatively little work in this area. The work on the effects of quantitative restrictions in imperfectly competitive markets is linked to the above question,[6] but to date, little work on what this might suggest about the price of a license seems to exist.

In this paper, I develop a series of models of monopoly that address this issue. The models show that the way in which licenses are sold, the demand conditions, and the market structure all influence the resulting price of a license. The results indicate that there is reason to expect that the price of a license may be much lower than that indicated by applying models of perfect competition. Thus, estimates of potential revenues such as those of the CBO and Bergsten et al. may be far too large. Moreover, if no revenues are to be raised from auctioning quotas unless they are very restrictive, the profit-shifting effect of such quotas, even when auctioned off, is unlikely to outweigh the loss in consumer surplus of such policies. For this reason, they are likely to have adverse welfare consequences even when set optimally.

I do *not* argue that in the real world license sales will raise no revenues. In the presence of uncertain demand they will, as licenses have an option value in this case. I merely point out that there is reason to expect revenues to be lower than those estimated under the assumption of perfect competition. I make my arguments in the simplest model, one without uncertainty. The uncertainty case is discussed in Krishna (1988b).

If there is a single foreign supplier of the product and markets are segmented, the price of a license is clearly zero. It is optimal for the monopolist to raise his price in response to a quota or VER so that the price of a license becomes zero. This model with segmented markets is mentioned in Hirofumi Shibata (1968), and it is developed diagrammatically in Takacs (1987) and in Paul Krugman and Elhanan Helpman (1989), who also look at foreign oligopoly.

However, one would expect that the presence of other markets and the possibility of arbitrage between them would make it optimal for the foreign monopolist to limit his price increase in response to a quota, thereby creating a positive price for the license. Thus, one might expect nonzero prices for licenses when markets are not segmented even with a foreign monopoly. Somewhat surprisingly, this is not necessarily so. Quotas set close to the free trade level *always* have a zero price. The only effect is an increase in the world price! This is the subject of Section I.

A simple example is developed in Section II in order to show how restrictive the quota has to be for the license price to become positive. The way that this varies with the relative size of the markets and demand elasticities is considered in addition to the welfare consequences of such policies. Section III briefly discusses the effect of an alternative timing structure on the results and argues that the spirit of the results

[5] Auction quota revenues in autos for 1987 are estimated at about $2.2 billion by Bergsten et al. (1989). The quota revenues fluctuate quite considerably over the years as demand fluctuates which alters the restrictiveness of the quota.

[6] See Krishna (1989b) for a survey of this work.

remains valid. Section IV contains some concluding remarks and directions for future research.

I. Foreign Monopoly with Costless Arbitrage

In this section it is assumed that there is a foreign monopolist who cannot price discriminate between his markets.[7] Let $Q(P)$ and $q(p)$ denote the demand functions facing the foreign firm in the home market and in the other market(s), respectively. Let $C[q+Q]$ denote its cost function.[8]

Let $R(P)$ be profits from sales in the home market. Assume that $R(P)$ is concave in P and is maximized at P^*. Similarly, let $r(p)$ be the profits from sales in other market(s), and let $r(p)$ be concave and maximized at p^*. It is easy to see that

$$P^*=\frac{\varepsilon}{\varepsilon-1}C \quad \text{and} \quad p^*=\frac{e}{e-1}C,$$

where ε and e are the respective demand elasticities, so that the monopolist in the absence of arbitrage would choose to charge a higher price in the market with less elastic demand. Because of arbitrage, however, the monopolist will choose one price that will be between the two prices he would have set in the absence of arbitrage possibilities. The optimal price for him to set maximizes $\pi(P)=R(P)+r(P)$, and is given by

$$P^M=\frac{\bar{\varepsilon}}{\bar{\varepsilon}-1}C,$$

where $\bar{\varepsilon}=\theta\varepsilon+(1-\theta)e$ and $\theta=Q/(q+Q)$. This is the free trade price. Thus, the monopolist chooses price as if he were faced with one market where the elasticity of demand is a share-weighted combination of the elasticities of the two markets. The question, then, is how a quota affects the price charged by the monopolist when the quota licenses are auctioned off.

At this point it is important to be clear about exactly what constitutes a license, how licenses are sold, and what the timing of moves is. With market segmentation, a license is defined to be a piece of paper that entitles its possessor to buy one unit of the product in question at the price charged in his market. If arbitrage is possible, then the possessor buys at the lower of the prices charged by the seller in the home and the world market. However, it is a dominated strategy for the monopolist to attempt to charge different prices in his different markets as sales will only be made at the lower of the two prices. For this reason, the monopolist can be restricted to choosing only *one* price.

Licenses are sold in a competitive market either to competitive domestic retailers with zero marginal costs of retailing or directly to consumers. In Sections I and II, I assume that the timing of moves is as follows. First, the government sets the quota. Then the monopolist sets his price. Finally, the market for licenses clears. This timing is consistent with the idea that the market for licenses clears *more* frequently than the monopolist sets prices, and that the government sets the quota even less frequently than the monopolist sets prices. Section III studies the case where the monopolist can adjust prices faster than the rate at which the market for licenses clears.

The model is then solved backwards as usual. First consider the market for licenses. If the price charged by the monopolist is P and the price of a license is L, then the demand for licenses must equal the demand for the good at price $P+L$, $Q(P+L)$. The supply of licenses is V, the level of the quota. The equilibrium price of a license is given by $L(P,V)$. $L(\cdot)$ is defined by the market for licenses clearing, so that $Q(P+L)=V$. Notice that if $Q(P)<V$, then $L(P,V)<0$ as defined thus far. However, since a quota is not binding if such a high price is charged, $L(P,V)$ is defined to be zero in this case. Let $P^V(V)$ be defined by $Q(P^V(V))=V$ so that $L(P,V)>0$ and the quota is binding if $P\leq P^V(V)$. By the defi-

[7] There may be domestic competitive suppliers in which case the monopolist's demand in what follows should be interpreted as the residual demand curve.

[8] Here marginal costs are assumed constant. Similar results obtain when marginal costs are not assumed constant.

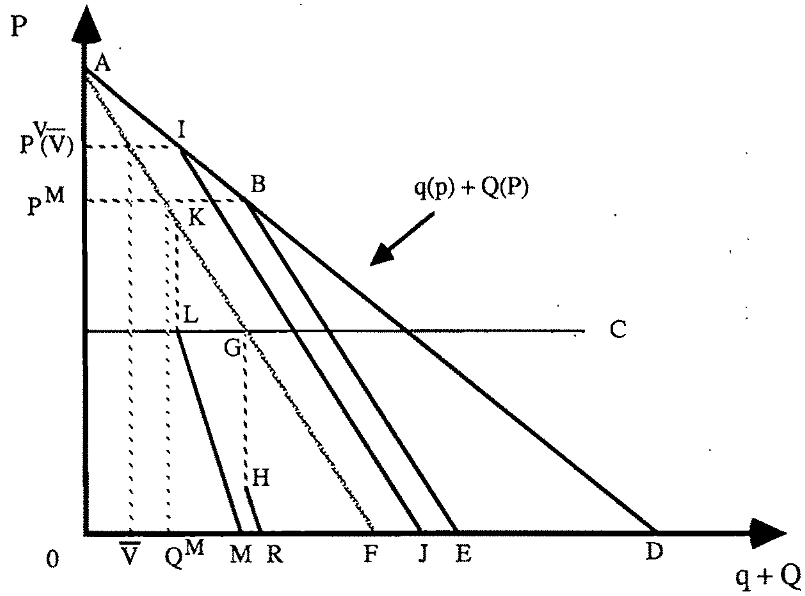


FIGURE 1

nition of $L(\cdot)$, it is apparent that if $P < P^V(V)$, then demand at home equals V, although V is less than $Q(P)$.

Now consider the total demand facing the monopolist with a quota. As demand in the home market has shrunk for $P < P^V(V)$, the total demand curve has a kink in it at $P^V(V)$. This is depicted in Figure 1 for the quota level $\bar{V}$.

In the absence of any quota, demand is given by AD, and marginal revenue by AF. The monopolist sets price equal to P^M, and sells $Q(P^M) + q(P^M)$. For convenience, Figure 1 is drawn so that $Q(P)$ and $q(P)$ are identical and linear. Hence the marginal revenue corresponding to total demand coincides with $Q(P)$. A quota at the free trade level, $Q(P^M)$, makes the demand facing the monopolist into ABE. This creates a kink in the demand curve at P^M. Marginal revenue is given by $AGHR$. Therefore, it remains optimal for the monopolist to price at P^M.

Now consider the effect of reducing the quota from $Q(P^M)$ to $\bar{V}$. $\bar{V}$ is the largest quota such that the intersection of marginal revenue and marginal cost occurs on the steeper but not vertical segment of the marginal revenue curve. This raises the price at which the quota binds to $P^V(\bar{V})$ from P^M, and the kink in demand occurs at $P^V(\bar{V})$. The demand curve with a quota at $\bar{V}$ is given by AIJ and the corresponding marginal revenue curve by $AKLM$. Notice that if V is close to Q^M, the profit maximizing point *must* occur at the intersection of the vertical part of the marginal revenue curve and marginal cost, C.[9] Therefore, the monopolist will find it optimal to charge $P^V(V)$ so that the price of a license is zero! Only if V is so small that the intersection of the marginal revenue curve (with a quota) and the marginal cost curve occurs on the steeper but *not* vertical segment of the marginal revenue curve will the price of a license be positive. This can only occur if V is substantially below Q^M. In Figure 1, any V lower than $\bar{V}$, the quota level depicted, gives $L(\cdot) > 0$.

[9] This is because when $V = Q^M$, the intersection of HR, the marginal revenue curve corresponding to BE, part of which is given by HR, with marginal cost would occur at a lower output level than the output level at which the kink in demand occurs. As the quota falls, the kink in the demand curve moves back twice as fast as the intersection of the marginal revenue curve with marginal cost. Finally, at the quota level of $\bar{V}$ depicted, the two coincide.

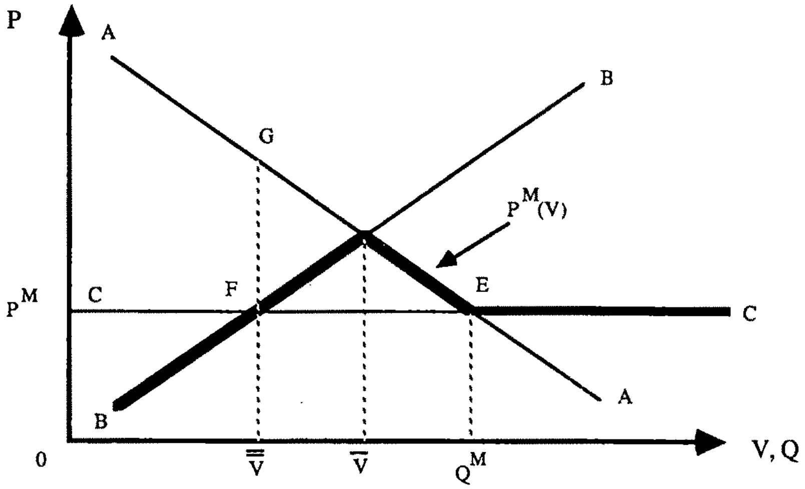


FIGURE 2

Figure 2 illustrates the effects of varying a quota. The domestic demand curve is depicted by AA. BB depicts the profit-maximizing price charged in both markets by the monopolist *assuming* that home demand is fixed at V independently of that price. As $r(P)$ is concave, BB is upward sloping. Note that when V, which along with Q is on the horizontal axis, equals Q^M, the price along BB must exceed P^M. Also, AA and BB intersect at output level $\bar{V}$ by the definition of $\bar{V}$ given previously.

The price charged when the quota is V is defined by $P^M(V)$, which is the dark line in Figure 2. If the quota is set above Q^M, it is not binding, so that the price charged lies along $\underline{EC}$. If the quota lies between Q^M and $\bar{V}$, so that it is "not too restrictive," the intersection of marginal revenue and cost occurs along the vertical segment of marginal revenue, so that the profit-maximizing price equals $P^V(V)$ and lies along AA. If the quota is below $\bar{V}$, so that it is "very restrictive," then the profit-maximizing price lies along BB. Therefore, the price charged by the monopolist first rises and then falls as the quota is reduced from the free trade level.

The license price is the vertical distance between $P^M(V)$ and the demand curve, AA. Only when $P^M(V)$ falls below AA can $L(\cdot)$, the license price, be positive. Since it is only when $L(\cdot)>0$ that auctioning quota licenses can raise revenue, it is only when the quota is set below $\bar{V}$ that welfare can rise. Moreover, since welfare falls as the quota falls from Q^M to $\bar{V}$, an even stronger condition is required for auctioning quotas to increase welfare relative to free trade. When V is set below $\bar{V}$, it is apparent from Figure 2 that the sum of consumer surplus and license revenues must fall short of free trade consumer surplus as long as $V \geq \bar{\bar{V}}$. Even at $V=\bar{\bar{V}}$, it is short by GFE, the deadweight loss of such a policy. Thus, only if $P^M(V)<P^M$, can welfare possibly rise.

This leaves the issue of auctioning existing quotas. Here the desirability of this policy depends on the existing method of implementation. If the VER is implemented by giving away licenses to foreign agents *other* than the foreign monopolist and the licenses are transferable in a competitive market, the monopolist's problem is the same whether or not licenses are auctioned. In this case, the only effect of auctioning licenses is to transfer license revenues, if they exist, from foreign license holders to the domestic government. In this case, auctioning licenses is better than giving them away only when the license price is positive, that is, when $V<\bar{V}$. This gives us Proposition 1, which addresses the two policy questions raised earlier.

PROPOSITION 1: *With a foreign monopoly and perfect and costless arbitrage, quotas at or close to the free trade level implemented by auctioning licenses yield no revenues and must reduce welfare. Quotas must be very restrictive if they are to raise welfare over free trade. Moreover, auction quotas dominate VERs at the same level only if the restriction is set at quite a low level. Note also that if the world price rises, a quota by one country reduces the welfare of other importing countries as well as that of the exporting country whose firm's profits fall with a quota.*

Other methods of implementing a VER would lead to a different status quo and so a different answer about the desirability of auctioning existing quotas. For example, if the VER were implemented by giving *all* the licenses to the *foreign* monopolist, then the foreign monopolist would take license revenues to be a part of his profits and this would affect his pricing behavior. This case is analyzed in Krishna (1989a).

Three further questions naturally arise. First, how restrictive must the quota be before a license commands a positive price? Second, how does the answer to this question depend on demand conditions? And finally, under what conditions can a quota that is auctioned off raise welfare over free trade? Next, a simple example is worked out that sheds some light on these questions.

II. An Illustrative Example

This example focuses on the role of the home market size relative to that of the other market and of demand elasticity in determining the effects of quota auctions.

It is assumed that consumers at home and abroad have identical constant elasticity demand functions given by $P^{-\varepsilon}$. There are, however, N consumers at home and n consumers abroad so that demand in the home market is $Q(P)=NP^{-\varepsilon}$ and that in the foreign market is $q(P)=nP^{-\varepsilon}$. As before, marginal costs are constant at C. Profit maximization in the absence of any quotas results in the monopolist charging $P^M=C\varepsilon/(\varepsilon-1)$, and selling $Q^M=N(C\varepsilon/(\varepsilon-1))^{-\varepsilon}$ at home. As usual, it is assumed that $\varepsilon>1$ so that profits are well behaved.

The smallest quota for which the license price is zero is depicted in Figure 1 by $\bar{V}$. When $\bar{V}$ is set as the quota, the marginal revenue curve associated with the market demand curve when the quota is binding intersects the marginal cost curve at exactly the level of the quota. If the quota is set at $\bar{V}$, the marginal revenue corresponding to the market demand curve when the quota is binding equals the marginal cost curve so that

$$\frac{d[(P-C)V+(P-C)nP^{-\varepsilon}]}{dP}=0.$$

This implies that

$$V-(P-C)\varepsilon nP^{-(\varepsilon+1)}+nP^{-\varepsilon}=0. \tag{1}$$

Moreover, for this to hold at the point where the constraint just binds, it must also be that

$$NP^{-\varepsilon}=V. \tag{2}$$

Substituting (2) into (1) gives

$$1-(P-C)\varepsilon\frac{n}{NP}+\frac{n}{N}=0$$

so that the price at which this occurs, $P^V(\bar{V})$, is given by

$$P^V(\bar{V})=\frac{C\varepsilon}{\varepsilon-1-\dfrac{N}{n}}. \tag{3}$$

Hence,

$$\bar{V}=N\left[\frac{C\varepsilon}{\left(\varepsilon-1-\dfrac{N}{n}\right)}\right]^{-\varepsilon}$$

and

$$(4)\qquad \frac{Q^M}{\bar{V}}=\left[\frac{(\varepsilon-1)}{\left(\varepsilon-1-\dfrac{N}{n}\right)}\right]^{\varepsilon}.$$

There are a few things to notice about this expression. First, Q^M always exceeds $\bar{V}$ so that the quota must be set below the free trade level of imports for a license price to be nonzero. Second, if $n=N$,

$$\frac{Q^M}{\bar{V}}=\left(\frac{\varepsilon-1}{\varepsilon-2}\right)^{\varepsilon}.$$

If $\varepsilon>2$, say, $\varepsilon=3$, then $Q^M/\bar{V}=8$ so that the quota needs to be quite restrictive for a license price to be positive. Third, for any ε and N, the limit as $n\to\infty$ of $Q^M/\bar{V}$ is 1. Thus, as the home market becomes small relative to the foreign one, the license price becomes positive when the quota is not very restrictive. Fourth, for any ε, as long as $\varepsilon-1-(N/n)>0$, $Q^M/\bar{V}$ rises as N/n rises. Fifth, if $\varepsilon-1-(N/n)<0$, the price of a license is always zero. As $\varepsilon-1-(N/n)$ approaches zero from above, $P^V(\bar{V})\to\infty$ and $\bar{V}\to 0$. Thus, the range where $L(\cdot)>0$, that is, where $V\leq\bar{V}$, shrinks to zero.

If e and ε can differ in the constant elasticity case, then the analogue of (1) implies that $P^M(V)>Ce/(e-1)$ as long as $V>0$. Since the free trade price, P^M, is $C\bar{\varepsilon}/(\bar{\varepsilon}-1)$, where $\bar{\varepsilon}=\theta\varepsilon+(1-\theta)e$ and θ is the share of the home market, P^M lies between $Ce/(e-1)$ and $C\varepsilon/(\varepsilon-1)$. Moreover, as $Ce/(e-1)$ rises as e falls, $P^M(V)$ *must* exceed P^M if $e\leq\varepsilon$. In this case, auction quotas must reduce welfare. If $e>\varepsilon$, $P^M(V)$ can lie below P^M and it is possible for welfare to rise when the optimal quota is set. The intuition for this is that if foreign elasticity of demand, e, is greater than home elasticity of demand, ε, then raising price to make the quota bind has a high cost in terms of losing customers in the foreign market. This is especially so if the home market is small relative to the rest of the world. It is possible to construct examples of cases where the optimal policy is a quota below the free trade level of imports.[10]

The results of this section are summarized in Proposition 2.

PROPOSITION 2: *If the elasticity of demand is constant and equal in both markets, then welfare* must *fall if auction quotas are imposed. If the home market demand is less elastic than that of the rest of the world, that is, $e>\varepsilon$, welfare* can *rise if the optimal quota is auctioned off. If home market demand is* more *elastic than the rest of the world, that is, $e<\varepsilon$, then welfare* must *fall.*

III. The Importance of Timing

So far we have assumed that the market for licenses clears faster than the monopolist sets his price. Thus, the monopolist can act like a Stackelberg leader and take into account the effect of his actions on the equilibrium price of a license. One might ask whether this timing structure is responsible for the results. Here I argue that this is not the case.

Consider the model of Section I with the new timing structure. In the last stage, the firm chooses price P taking as *given* the value of L and V. Its profits thus depend on how consumer demand is affected given this level of L and V. If consumers assume that L is fixed and that any number of licenses will be available at this price, their demand for the good is given by $Q(P+L)$ even if P is very low. I call this the case with myopic consumers. In this case there is no kink in the demand curve facing the monopolist. The equilibrium license price, $L(V)$, can be shown to be exactly the specific tariff that leads to V being demanded. This gives Proposition 3.

PROPOSITION 3: *When firms take L as given, the equilibrium price of a license is zero if $V\geq V^F$. It is positive if $V<V^F$ and increases as V decreases. Thus, for $V<V^F$, a license always has a positive price.*

[10] Assuming $\varepsilon=1.1$, $e=8$, $N/n=0.0001$ yields one such example.

If, on the other hand, consumers do not assume that L is fixed even if P is very low, we have the case with nonmyopic consumers. Here consumers realize that the number of licenses is limited to V. They infer that if the monopolist charges a very low price so that $Q(P+L)>V$, that is, $P<P^V(V)-L$, then the shadow price of a license will exceed L and equal $\tilde{L}$ where $Q(P+\tilde{L})=V$, so that $\tilde{L}=P^V(V)-P$. This will give the monopolist a total demand of $q(P)+V$ instead of $q(P)+Q(P+L)$.

Consumers therefore take the license price as given when the product price is high, but they realize that a low product price creates a black market for licenses and raises the effective license price. This asymmetry can be shown to create a continuum of equilibrium license prices. A zero license price remains an equilibrium as long as the quota is not too restrictive. The results are summarized in Proposition 4.

PROPOSITION 4: *With nonmyopic consumers, zero remains an equilibrium license price as long as the quota is not too restrictive. For more restrictive quotas, equilibrium license prices are bounded away from zero. However, there is a continuum of such license prices for any quota below the free trade level.*[11]

Thus, the result that auction quotas may not raise revenues unless they are quite restrictive reemerges even when the timing of moves is altered and consumers are not myopic. However, it is less compelling here as other equilibria with positive license prices also exist.

IV. Conclusion

The main thrust of this paper is that with foreign monopoly the effects of auctioning quota licenses must take into account the induced effect on the pricing policy of the monopolist. Doing so casts doubt on the revenue-raising potential of such policies and this adversely affects their welfare consequences. Though this paper deals with foreign monopoly, the point is more general. Similar results go through with foreign oligopoly as well, as shown in Krishna (1988a).

Still, much remains to be done to determine the desirability of auction quotas. First, their desirability under uncertainty needs more study. Second, it may be possible to use recent work on computable partial equilibrium models, such as that of Avinash Dixit (1987) and Anthony Venables and Alasdair Smith (1986), to help build empirically implementable models to give estimates of the welfare effects of auctioning quota rights in particular markets. Third, while this paper assumes the market for licenses is competitive, it would be desirable to study the determinants of the market structure in the market for licenses. Fourth, as the details of the implementation procedure, for both VERs and auctioning quotas are crucial, more attention to these is warranted. I am currently working along these lines.

SYMBOLIC NOTATION

P	Price in the home market
p	Price in the foreign market
$Q(P)$	Demand in the home market
$q(p)$	Demand in the foreign market
$C[q+Q]$	Cost function
$R(P)$	Profits in the home market
$r(p)$	Profits in the foreign market
ε	Demand elasticity in the home market
e	Demand elasticity in the foreign market
θ	Share of the home market in sales $=\dfrac{Q}{q+Q}$
$\bar{\varepsilon}$	Share weighted average of elasticities $\theta\varepsilon+(1-\theta)e$
V	Quota level
P^M	Price under free trade
$P^V(V)$	Price at which the quota just binds
$P^M(V)$	Price charged by monopolist when quota is V
$L(P,V)$	License price as a function of P and V
P^*	Price at which $R(P)$ is maximized
p^*	Price at which $r(p)$ is maximized
$\pi(P)$	Total profits which equal $R(P)+r(P)$
$\bar{V}$	Smallest quota where $L(P,V)$ equals zero
$\bar{\bar{V}}$	Quota where $P^M(V)=P^M$
N	Size of the home market
n	Size of the foreign market
$\tilde{L}$	Effective license price with nonmyopic consumers
Q^M	Free trade imports

[11] For detailed proofs of Propositions 3 and 4, see Krishna (1989b).

REFERENCES

Bergsten, C. Fred et al., "Auction Quotas and United States Trade Policy," *Policy Analyses in International Economics*, Institute for International Economics, Washington, September 1987, *19*.

Dixit, Avinash, "Tariffs and Subsidies Under Oligopoly: The Case of the U.S. Automobile Industry," in H. Kierzkowski, ed., *Protection and Competition in International Trade*, Oxford and New York: Basil Blackwell, 1987.

Hufbauer, Gary and Rosen, Howard F., "Trade Policy for Troubled Industries," *Policy Analyses in International Economics*, Institute for International Economics, Washington, 1986, *15*.

Krishna, Kala, (1988a) "The Case of the Vanishing Revenues: Auction Quotas with Oligopoly," NBER Working Paper No. 2723, 1988.

______, (1988b) "Auction Quotas with Uncertainty," Harvard University, unpublished manuscript, 1988.

______, (1989a) "Making Altruism Pay in Auction Quotas," Harvard University, unpublished manuscript, 1989.

______, (1989b) "The Case of the Vanishing Revenues: Auction Quotas with Monopoly," NBER Working Paper No. 2890, 1989.

______, (1989c) "What Do VERs Do?" in Ryuzo Sato and Julianne Nelson, eds., *U.S.-Japan Trade Relations*, Cambridge: Cambridge University Press, 1989.

Krugman, Paul and Helpman, Elhanan, *Market Structure and Trade Policy*, Cambridge, MA: MIT Press, 1989.

Lawrence, Robert Z. and Litan, Robert E., "Saving Free Trade," Washington, Brookings Institution, 1986.

Shibata, Hirofumi, "A Note on the Equivalence of Tariffs and Quotas," *American Economic Review*, March 1968, *68*, 137–46.

Takacs, Wendy E., "Auctioning Import Quota Licenses: An Economic Analysis," Institute for International Economic Studies, University of Stockholm, Seminar Paper No. 390, September 1987.

Tarr, David, *A General Equilibrium Analysis of the Welfare and Employment Effects of U.S. Quotas in Textiles, Autos and Steel*, Bureau of Economics, Staff Report to the Federal Trade Commission, February 1989.

Venables, Anthony and Smith, Alasdair, "Trade and Industrial Policy Under Imperfect Competition," *Economic Policy*, October 1986, *3*, 622–71.

Tying, Foreclosure, and Exclusion

By Michael D. Whinston*

In recent years, the "leverage theory" of tied good sales has faced heavy and influential criticism. In an important sense, though, the models used by its critics are actually incapable of addressing the leverage theory's central concerns. Here I reconsider the leverage hypothesis and argue that tying can indeed serve as a mechanism for leveraging market power. The mechanism through which this leverage occurs, its profitability, and its welfare implications are discussed in detail. (JEL *610)*

A firm engages in tying when it makes the sale (or price) of one of its products conditional upon the purchaser also buying some other product from it. Tying has a long history of scrutiny under the antitrust laws of the United States, and throughout this history it has been harshly treated by the courts.[1] A primary basis for this condemnation has been the courts' belief in what has come to be known as the "leverage theory" of tying: that is, that tying provides a mechanism whereby a firm with monopoly power in one market can use the leverage provided by this power to foreclose sales in, and thereby monopolize, a second market.

In recent years the leverage theory has come under heavy attack from a number of authors whose arguments are traceable to the University of Chicago oral tradition associated with Aaron Director (see, for example, Director and Edward Levi, 1956; Ward S. Bowman, 1957; Richard A. Posner, 1976; and Robert H. Bork, 1978). A typical rendition of their criticism goes along the following lines: Suppose that a firm is a monopolist of some good A that a consumer values at level v_A and that costs c_A to produce. The consumer also consumes some other competitively supplied product B that she values at level v_B and that can be produced at a unit cost of c_B. Now, the monopolist *could* require the consumer to purchase good B from him if she wants good A, but what will he gain? The consumer will only purchase such a bundle if its price is no larger than $v_A + c_B$, and so the monopolist can do no better than earning $(v_A - c_A)$, the level he earns selling good A independently. In short, there is only one monopoly profit that can be extracted.

Similar arguments are given for the case of complementary products. Richard Posner (1976), for example, comments as follows:

> [A fatal] weakness of the leverage theory is its inability to explain *why* a firm with a monopoly of one product would want to monopolize complementary products as well. It may seem obvious..., but since the products are by

*Department of Economics, Harvard University, Cambridge, MA 02138. Much of the work on this paper was done during the summers of 1986 and 1987. I would like to thank the Federal Trade Commission and the Graduate School of Business at Stanford University for their hospitality during that time. I would also like to thank Patrick Bolton, Franklin Fisher, Louis Kaplow, Richard Schmalensee, and Jean Tirole for their helpful comments and George and Elizabeth Whinston for their editing assistance. Financial support from the National Science Foundation through grant no. SES-8618775 is also gratefully acknowledged.

[1] Tying doctrine was originally developed in patent cases (*Motion Pictures Patents Co. v. Universal Film Manufacturing Co.*, 243 U.S. 502 (1917)). Since then a long line of case law has developed under both Section 1 of the Sherman Act and Section 3 of the Clayton Act. (See, for example, *International Salt v. U.S.*, 332 U.S. 392 (1947) and *Northern Pacific Railway Co. v. U.S.*, 356 U.S. 1 (1958).) Similar ideas have also been developed under Section 2 of the Sherman Act. (See, for example, *U.S. v. Griffith*, 334 U.S. 100 (1948) and *U.S. v. United Shoe Machinery Corp.*, 110 F. Supp. 295 (D. Mass. 1953).) Two cases involving less harsh treatment are *Times-Picayune Publishing Co. v. U.S.*, 345 U.S. 594 (1953) and *U.S. v. Jerrold Electronics Corp.*, 365 U.S. 567 (1961).

> hypothesis used in conjunction with one another..., it is not obvious at all. If the price of the tied product is higher than the purchaser would have to pay on the open market, the difference will represent an increase in the price of the final product or service to him, and he will demand less of it, and will therefore buy less of the tying product. To illustrate, let a purchaser of data processing be willing to pay up to $1 per unit of computation, requiring the use of 1 second of machine time and 10 punch cards, each of which costs 10 cents to produce. The computer monopolist can rent the computer for 90 cents a second and allow the user to buy cards on the open market for 1 cent, or, if tying is permitted, he can require the user to buy cards from him at 10 cents a card —but in that case he must reduce his machine rental charge to nothing, so what has he gained? [p. 173]

Thus, the critics contend, if a monopolist does employ tying, his motivation cannot be leverage. In its place, they point to a number of socially beneficial, or at worst ambiguous, alternative explanations for tying: for example, price discrimination (Bowman, 1957), achieving economies of joint sales, protection of goodwill, risk sharing, and cheating on a cartel price. Almost inadvertently, the more formal economics literature on tying (Meyer L. Burstein, 1960; Roger D. Blair and David L. Kaserman, 1978; Richard Schmalensee, 1982) has reinforced this view as a result of its exclusive focus on price discrimination motivations for the practice. Thus, Posner (1976) goes on to note that "the replacement of leverage by price discrimination in the theory of tie-ins has been part of the economic literature for almost twenty years."[2] These criticisms have, in fact, had a tremendous impact in both legal and economic circles.[3]

In an important sense, however, the existing literature does not really address the central concern inherent in the leverage theory, namely, that tying may be an effective (and profitable) means for a monopolist to affect the market structure of the tied good market (i.e., "monopolize" it) by making continued operation unprofitable for tied good rivals. The reason lies in the literature's pervasive (and sometimes implicit) assumption that the tied good market has a competitive, constant returns-to-scale structure. With this assumption, the use of leverage to affect the market structure of the tied good market is actually impossible. Thus, in contrast to a concern over the effects of tying on market structure, the existing literature's focus is on a demand-side notion of "leverage": the idea that, taking the prices charged by tied good competitors as given, a firm might be able to extract greater profits from consumers by tying.[4]

In this paper, I reexamine the leverage hypothesis. In particular, I examine several simple models that depart from the competitive, constant returns-to-scale structure assumed in the existing literature. In contrast,

[2] Bork (1978) sums up his discussion of tying more emphatically: "[The leverage] theory of tying arrangements is merely another example of the discredited transfer of power theory, and perhaps no other variety of that theory has been so thoroughly and repeatedly demolished in the legal and economic literature."

[3] In a recent antitrust textbook, for example, Blair and Kaserman (1985) comment that "according to this view, somehow the seller expands or levers his monopoly power from one market to another. This, of course, is not possible. A seller cannot get two monopoly profits from one monopoly.... Thus, the leverage theory of tying is unsatisfactory." The 1985 Department of Justice Vertical Restraints Guidelines state that "Tying arrangements often serve procompetitive or competitively neutral purposes.... [They] generally do not have a significant anticompetitive potential." For a recent rebuttal to this view in the legal literature, however, see Louis Kaplow (1985).

[4] Indeed, this is exactly the sense in which the existing literature can be said to focus on price discrimination aspects of the practice; it analyzes whether tying is a profitable strategy *given* the prices of tied good competitors (which can be thought of as creating an induced demand structure for the monopolist). In contrast, here my focus is on the ability of tying to *change* those prices, in particular, by making continued operation unprofitable for competitors.

here I assume that scale economies exist in the production process for the tied good, and as a result, the structure of that market is oligopolistic.

In these models I address three basic questions. First, can tying succeed in altering the market structure of the tied good market, and if so, how? Second, is it a profitable strategy? Third, what are the welfare consequences? As we shall see, tying can lead to a monopolization of the tied good market. Most interestingly, the mechanism through which this exclusion occurs is foreclosure; by tying, the monopolist reduces the sales of its tied good market competitor, thereby lowering his profits below the level that would justify continued operation.

Tying is frequently a profitable strategy for the monopolist in these models, and it is often so precisely because of its potential for altering the market structure of the tied good market. The particular circumstances in which tying is a desirable strategy for the monopolist, however, depend in part on whether he is able to make a precommitment to tie. In many circumstances this is indeed possible. One of the primary ways in which this can be accomplished is through product design and the setting of production processes, both of which may involve significant sunk costs. By bundling components of its system together or by making interfaces between the separately sold components incompatible with their rivals' components, firms can precommit to their marketing strategy. IBM, for example, was accused of incorporating increased amounts of storage into its central processing units in order to prevent sales by plug compatible memory manufacturers and also of trying to achieve interface incompatibility for the same purpose (Franklin M. Fisher, John J. McGowan, and Joen E. Greenwood, 1983, pp. 332–33). Kodak was accused of designing its new film and camera in a format incompatible with rival manufacturers' products (*Berkey Photo v. Eastman Kodak Co.*, 603 F. 2d 263 (2d Cir., 1979)).

On the other hand, in a significant number of tying cases little more than an easily changed marketing decision seems to be involved. For example, in *Times-Picayune Publishing Co. v. U.S.* (345 U.S. 594 (1953)), the publisher of the only morning newspaper in New Orleans only sold an advertisement in his morning paper with an advertisement in that day's evening newspaper (which faced competition from another evening newspaper). In *U.S. v. Griffith* (334 U.S. 100 (1948)), a movie theater chain refused to show films in its theaters in towns in which it possessed a monopoly if the distributor did not give it that film in towns where it faced competition. In *United Shoe Machinery Corp. v. U.S.* (110 F. Supp. 295 (D. Mass. 1953)), United Shoe bundled repair service with its shoe machinery leases.

Finally, when tying does lead to exclusion of rivals, the welfare effects both for consumers and for aggregate efficiency are in general ambiguous. The loss for consumers arises because, when tied market rivals exit, prices may rise and the level of variety available in the market necessarily falls. Indeed, in the models studied here, tying that leads to the exit of the monopolist's tied market rival frequently leads to increases in all prices, making consumers uniformly worse off. More generally, though, as is common in models of price discrimination, some consumers may be made better off by the introduction of tying. The effect on aggregate welfare, on the other hand, is uncertain because of both the ambiguous effects of price discrimination and the usual inefficiences in the number of firms entering an industry in the presence of scale economies and oligopolistic pricing (A. Michael Spence, 1976; N. Gregory Mankiw and Whinston, 1984).

Though most tying cases involve products that are complements (particularly those where precommitment is involved), for expositional purposes I begin below by considering the case of independent products. In Section I, I first analyze the simple case where all consumers have an identical valuation of the monopolized product, so that the monopolist, if he chooses to price his goods independently, can fully extract all of the surplus from his monopolized good. I show that, *absent precommitment*, tying is not a useful strategy for the monopolist; any equilibrium outcome will be equivalent to one where only independent pricing is allowed. Despite this fact, however, a *precommitment* to tying can be a profitable strategy for the monopolist because of its potential for excluding his tied market rival. This exclusionary effect arises because of what I call "strategic foreclosure": tying represents a commitment to foreclose sales in the tied good market, which can drive its rival's profits below the point where remaining in the market is profitable. This strategic incentive to foreclose sales in the tied good market occurs because once the monopolist has committed to offering only tied sales, it can only reap its profit from its monopolized product by making a significant number of sales of the tied good. Thus, in this model, tying necessarily lowers the profits of the monopolist's tied good rival. I then discuss the implications of such a commitment to tying for the monopolist's profits, for consumers, and for aggregate efficiency, and present a simple example to illustrate these points.

In Section II, I investigate how the presence of heterogeneous preferences among consumers for the monopolized good affects these results. Two basic findings emerge. First, with heterogeneous preferences for the tying good, tying no longer necessarily results in strategic foreclosure and the lowering of the monopolist's tied good rival's profits (though it still does in many circumstances). If, for example, a significant number of consumers in the tied market have low valuations of the tying good, tying (not surprisingly) will not be a successful exclusionary device. In addition, a more subtle effect may prevent a commitment to tying from lowering the tied good rival's profits. This occurs when tying substantially decreases the responsiveness of the monopolist's demand to price changes relative to the level previously prevailing in the tied good market.

Second, with heterogeneous valuations, tying can now also be a profitable strategy in the absence of precommitment. There are two senses in which this is true. First, in a purely static sense, the monopolist may find tying to be a profitable strategy given its rival's price. This motivation for tying is analogous to that in the monopolistic bundling literature (for example, W. J. Adams and J. L. Yellen, 1976; R. Preston McAfee, John McMillan, and Whinston, 1989), but here it can have important competitive effects: tied product rivals can find their sales foreclosed and continued operation unprofitable. Second, even when tying is not profitable in this static sense, it may be in a dynamic sense when the exclusion of rivals through predation is possible. In such cases, tying can be a profitable strategy for the monopolist precisely because it forecloses the sales of the monopolist's tied market rival.

In Section III, I turn to the case of complementary products used in fixed proportions. I first consider a model of fixed proportions that is essentially an extension of the simple example quoted above from Posner (1976) to the case where the tied good market involves scale economies and oligopolistic behavior. Despite these differences, Posner's central contention continues to hold: a monopolist of one component never finds it worthwhile to tie in order to reduce the level of competition in the market for the other component. The reason lies in the fact that when the monopolized product is essential for all uses of the two products, the monopolist can always benefit from more competition in the nonmonopolized market through sales of its monopolized product. Nevertheless, I then show that in two natural extensions of this model where the monopolized product is no longer essential for all uses of the nonmonopolized components, tying once again emerges as a profitable exclusionary strategy. In one case, the presence of an inferior, competitively supplied alternative to the monopolized component leads to results that parallel those for independent products. In the other case, the existence of a second use for the nonmonopolized product (such as a replacement part market) can give the monopolist an incentive to tie in order to reduce competition in this other market.

Finally, I conclude in Section IV with a brief discussion of the implications of these findings.

I. Independent Products

I begin by considering an extremely simple model with independent products. There are two markets, which I label A and B. Market A is monopolized by firm 1 (say, because of a patent). Market B, on the other hand, is potentially served by two firms, firm 1 and firm 2. The products of firms 1 and 2 in market B are differentiated. Production in market B involves fixed costs of K_i plus an expenditure of c_{Bi} per unit for firm i. Unit costs for good A are c_A. For expositional simplicity, I ignore the possibility that there are fixed costs for product A.

Consumers, who are indexed by $d \in (0,1)$ with total measure 1, each desire at most one unit of good A and one unit of good B. All consumers have a reservation value of $\gamma > c_A$ for good A, while a consumer of type d has a valuation of $v_{Bi}(d)$ for a unit of firm i's product B. Resale of products by consumers is assumed to be prohibitively costly. In the absence of tying by firm 1, consumers simply respond to individual product prices (P_A, P_{B1}, P_{B2}). Firm i's sales of product Bi are then given by some function $x^i(P_{B1}, P_{B2}) \leq 1$, which I assume to be everywhere differentiable and satisfy (subscripts denote partial derivatives) $x^i_j(P_{B1}, P_{B2}) \geq 0$ if $j \neq i$ and ≤ 0 if $j = i$, with strict inequalities if $x^i(\cdot,\cdot) \in (0,1)$. That is, products $B1$ and $B2$ compete with each other for consumer purchases.

When bundling is not permitted (which I will refer to below as an "independent pricing game"), it is easy to see that firm 1 will always set $P_A = \gamma$. It is also useful for what follows to define each firm i's best response correspondence in market B by $P^*_{Bi}(P_{Bj})$, which solves

$$\max_{P_{Bi}} (P_{Bi} - c_{Bi}) x^i(P_{B1}, P_{B2}).$$

I assume that this correspondence is single-valued, continuous, and differentiable with $P^{*\prime}_{Bi}(P_{Bj}) \in (0,1)$ (so products $B1$ and $B2$ are strategic complements in the sense of Jeremy I. Bulow, John D. Geanakoplos, and Paul D. Klemperer, 1985).

In the next two subsections I analyze the use of tying both for cases where firm 1 can precommit to tie and where it cannot. For the case without precommitment, I analyze a simple two-stage game. In stage one, each firm simultaneously decides whether to be active in market B. If firm i decides to be active, it incurs the cost K_i. In stage two, the firms pick prices (simultaneously if both are active). If firm 1 is active in market B, it can offer three different items for sale: good A at a price of P_A, good $B1$ at a price of P_{B1}, and a bundle consisting of one unit of good A and one unit of good $B1$ at a price of $\bar{P}$. If firm 2 is active, on the other hand, it can only offer good $B2$ at price P_{B2}. Throughout I assume that firm 1 is unable to monitor customer purchases; this assumption rules out the use of requirements contracts (where a consumer agrees as a condition of buying good A not to buy good $B2$) and also implies that a bundle will be purchased only if $\bar{P} \leq P_A + P_{B1}$.

To analyze the case where precommitment is possible, I extend this game to three stages. In the (new) first stage of the game, firm 1 commits to which subset of three possible products—good A, good $B1$, and a bundle—it will be able to produce. For example, firm 1 can commit itself to a position where it will only be able to produce a bundle. The second and third stages are then identical to the no commitment game, but with firm 1 only able to offer for sale those items that it is able to produce.[5] Thus, as discussed in the introduction, by setting its design and production process, firm 1 is able to commit to a tying strategy.

Finally, at various points below I make comparisons between the outcomes of these two games and those of a game where firm 1 only offers goods A and $B1$ indepen-

[5] Note that as long as firm 1 can produce both goods A and $B1$ separately it can still offer a bundle for sale.

dently (more precisely, a game that is the same as the no precommitment game but where bundling is prohibited). I refer to this game as the "independent pricing game." Firm 1 is said to tie whenever its pricing is not identical (or, more generally, economically equivalent) to that arising in this independent pricing game.

A. *Tying Without Precommitment*

Consider first the no commitment game. If firm 1 is active in market B, then in the second stage of this game it selects three (nonnegative) prices: $(P_A, P_{B1}, \bar{P})$. As the following proposition makes clear, however, tying is not a useful strategy in this game.

PROPOSITION 1: *Any subgame perfect equilibrium outcome of the no commitment game is economically equivalent to a subgame perfect equilibrium outcome in the independent pricing game.*

PROOF:

The proposition is established by arguing that in the subgames of the no commitment game in which firm 1 is active in market B, any Nash equilibrium in prices is equivalent to a Nash equilibrium in the corresponding subgame of the independent pricing game. Then, given the equivalence of the equilibria in the pricing subgames, firms' decisions about whether to be active in market B must also be equivalent in the two games.

Consider the subgame where both firms are active in market B. The equivalence of equilibria is demonstrated by arguing that for any set of prices $((P_A^0, P_{B1}^0, \bar{P}^0); P_{B2}^0)$ that constitute a Nash equilibrium in the no commitment game there is a set of independent prices $(\hat{P}_A, \hat{P}_{B1})$ such that sales and profits are the same for *both* firms under prices $((\hat{P}_A, \hat{P}_{B1}); P_{B2})$ as under prices $((P_A^0, P_{B1}^0, \bar{P}^0); P_{B2})$ when $P_{B2} = P_{B2}^0$ and are the same for firm 2 for *any* P_{B2}. This implies that $((\hat{P}_A, \hat{P}_{B1}); P_{B2}^0)$ is a Nash equilibrium in the independent pricing game (note that firm 1 now has fewer possible deviations).

The equivalence clearly holds if firm 1's equilibrium strategy has $\bar{P}^0 > P_A^0 + P_{B1}^0$, so suppose that $\bar{P}^0 \le P_A^0 + P_{B1}^0$. There are two cases to consider. First, suppose that $P_A^0 > \gamma$. If this is firm 1's best response, then it must be that all consumers are buying firm 1's bundle since otherwise firm 1 could do better by setting $P_A = \gamma$ while leaving all of its other prices unchanged: this price would make profitable sales of product A to those consumers not buying the bundle, while having no effect on firm 1's sales of either good $B1$ or the bundle (since consumers are indifferent about buying good A at this price).[6] In addition, since all consumers are purchasing the bundle (and therefore none are purchasing either A or $B1$ alone) it cannot be that $\bar{P}^0 < \gamma$ since, if it were, firm 1 could do better by offering only the bundle at a price of γ. But, if so, then setting $(\hat{P}_A = \gamma, \hat{P}_{B1} = \bar{P}^0 - \gamma)$ yields identical sales and profits to both firms given P_{B2}^0 and identical profits to firm 2 for all P_{B2}. Second, suppose instead that $\gamma \ge P_A^0$. Note first that we must have $\bar{P}^0 \ge P_A^0$ in such an equilibrium: otherwise all consumers would be buying firm 1's bundle (all consumers would be willing to buy good A individually and they can get good A cheaper by buying the bundle) and firm 1 would increase its profits by offering only the bundle at a price of γ. But if $\gamma \ge P_A^0$ and $\bar{P}^0 \ge P_A^0$, then each consumer buys either good A alone or the bundle from firm 1. In this case, prices of $(\hat{P}_A = P_A^0, \hat{P}_{B1} = \bar{P}^0 - P_A^0)$ yield identical sales and profits for both firms for all P_{B2}.

A similar argument establishes the equivalence for the subgame where only firm 1 is active. □

The basic idea behind Proposition 1 is fairly straightforward. First, it is always worthwhile for firm 1 to make sure that all consumers purchase product A either alone or in the bundle. Given that all consumers are consuming good A, however, if firm 1 engages in tying, then consumers choose between buying only good A or the bundle from firm 1. They do so by imputing

[6] I assume that all consumers will buy good A when $P_A = \gamma$. This assumption can be avoided through the use of limiting arguments, but it is made in order to ease the exposition.

an effective price of $(\bar{P}_1 - P_A)$ $((\bar{P}_1 - \gamma)$ if $P_A > \gamma)$ to the product $B1$ portion of the bundle, so that tying is effectively equivalent to an independent pricing strategy.

B. *Commitment and Strategic Foreclosure*

The negative result of Proposition 1 changes dramatically if firm 1 is able to precommit to tying through its choice of which goods it will be able to produce. In the three-stage game that I have described above, firm 1 can choose to produce seven different sets of goods: both goods individually, both goods individually and also a bundle, the bundle only, the bundle and product A, the bundle and product $B1$, A only, and $B1$ only. The argument in Proposition 1 implies that the first two options both yield outcomes equivalent to those in the independent pricing game and so they are strictly better for firm 1 than the last two (which yield lower profits to firm 1 in any subgame where it is active and at least as large profits to firm 2 when it is active). In fact, the following two lemmas indicate that firm 1's choice is essentially between producing independent goods and producing only the bundle.

LEMMA *1: Any subgame perfect equilibrium outcome in the subgame of the commitment game where firm* 1 *can produce only the bundle and product A is equivalent to a subgame perfect equilibrium outcome of the independent pricing game.*

PROOF:

The argument closely parallels that used to prove Proposition 1 and is omitted here. □

LEMMA 2: Any subgame perfect equilibrium outcome in the subgame of the commitment game where firm 1 *can only produce the bundle and product B*1 *is equivalent to a subgame perfect equilibrium outcome that arises in the subgame of the commitment game where firm* 1 *can only produce a bundle.*

PROOF:

In Appendix A. □

Given these results, firm 1 can restrict its attention to either producing goods A and $B1$ separately, which yields an outcome equivalent to that in the independent pricing game, or to committing to producing only a bundle. I now turn to an investigation of the competitive effects of this tying strategy. As the following result makes clear, such a commitment may make it unattractive for firm 2 to be active in the market.

PROPOSITION 2: *In the subgame of the commitment game where both firms are active and firm* 1 *has committed itself to producing only the bundle, firm* 2 *earns less than it does in the independent pricing game.*

PROOF:

In Appendix A. □

One might at first think that bundling in this context would have no effect at all: if firm 1 were charging independent prices of $P_A = \gamma$ and P_{B1}, a switch to bundling at a total price of $\gamma + P_{B1}$ would not change the demand for good $B1$ at all. The intuition for Proposition 2, however, centers on the way in which firm 1's pricing incentives change when it bundles. In an independent pricing game, firm 1's best response $P^*_{B1}(P_{B2})$ satisfies,

$$(1)\quad [P^*_{B1}(P_{B2}) - c_{B1}]\,x^1_1(P^*_{B1}(P_{B2}), P_{B2}) + x^1(P^*_{B1}(P_{B2}), P_{B2}) = 0.$$

By contrast, when firm 1 bundles and sets price $\bar{P}$, the demand for its bundle is given by $x^1(\bar{P} - \gamma, P_{B2})$ and its best response to firm 2's price P_{B2} by $\bar{P}^*(P_{B2})$ such that

$$(2)\quad [\bar{P}^*(P_{B2}) - c_A - c_{B1}] \times x^1_1(\bar{P}^*(P_{B2}) - \gamma, P_{B2}) + x^1(\bar{P}^*(P_{B2}) - \gamma, P_{B2}) = 0.$$

Note first that if $\gamma = c_A$, then $\bar{P}^*(P_{B2}) = P^*_{B1}(P_{B2}) + \gamma$. However, if $\gamma > c_A$, then at $\bar{P} = P^*_{B1}(P_{B2}) + \gamma$ the left-hand side of (2) is strictly negative. Thus, it must be that $\bar{P}^*(P_{B2}) < P^*_{B1}(P_{B2}) + \gamma$: firm 1's optimal ef-

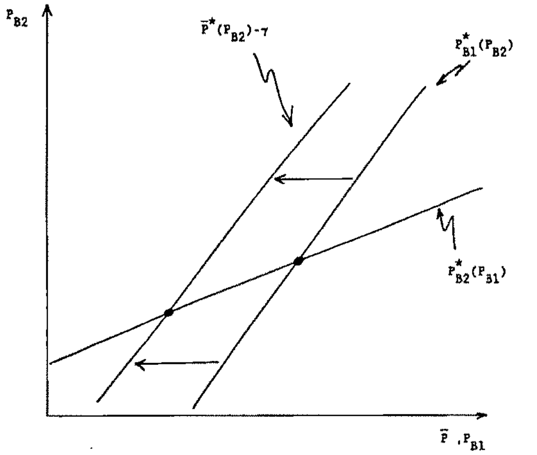


FIGURE 1

fective price for good $B1$ is lower under bundling than under independent good pricing. The reason is straightforward: when firm 1 is bundling, in order to make profitable sales of its monopolized product, good A, it must also make sales of good $B1$. This leads it to cut price in an effort to take sales away from firm 2, an effect I call "strategic foreclosure."[7] The effect on the equilibrium can be seen in Figure 1, where the equilibrium effective price for $B1$ and actual price for $B2$ both fall as a result of firm 1's bundling, thereby lowering firm 2's profits. Thus, by committing to tie by producing only a bundle, firm 1 may make continued operation unprofitable for its tied good rival. This point emerges particularly clearly in the following simple example, which is a limiting case of the above model.

Example 1. Suppose that all consumers view products $B1$ and $B2$ as perfect substitutes with value v, that $(c_{B1}-c_{B2})>K_2>0$ and, to focus attention on firm 2's activity decision, that $K_1=0$ (this could be a situation of entry deterrence where only firm 1 has already sunk its market B set-up costs). Then the subgame perfect equilibrium outcome of the independent pricing game has firm 2 being active in market B, making all sales in that market, and earning profits of $(c_{B1}-c_{E2})-K_2>0$.[8] By contrast, if $(c_{B2}-c_{B1})+(\gamma-c_A)>0$ and firm 1 commits to bundling, firm 2 earns zero if it is active, and so the unique equilibrium outcome involves firm 2 being inactive and firm 1 extracting all of the consumers' surplus. ■

Note that if both firms are active, firm 1's profits are also lower in the bundling regime than under independent pricing. This is true because bundling not only loses some profitable sales of good A, but also causes firm 2 to lower its price.[9] Thus, in this model, firm 1 would never commit to tying unless this would succeed in driving firm 2 out of the market.[10]

[7] Jean Tirole has pointed out a nice analogy to situations in which firms can invest in cost reduction. Here, by bundling, firm 1 can incur an "investment cost" of $(\gamma-c_A)$ (the lost good A sales) but thereby lowers its effective marginal cost in market B by $(\gamma-c_A)$. This lowering of marginal cost makes firm 2 more agressive in market B. As noted in Drew Fundenberg and Tirole (1984) (see also Tirole, 1988), with price competition (strategic complements) and entry deterrence/exit inducement, firms overinvest in cost reduction relative to what they would do absent this strategic effect (a "top dog" strategy), a comparison analogous to my commitment versus no commitment games.

[8] I am ignoring subgame perfect equilibria in which a firm prices below cost and makes no sales. These equilibria involve weakly dominated strategies and can be formally eliminated here through the use of R. Selten's (1975) notion of trembling-hand perfect equilibria (formally one examines discrete approximations to the game considered in the text where prices must be named in some discrete unit of account).

[9] Note that these lower profits can potentially force firm 1 to exit should it commit to tying and have positive fixed costs in market B (if it believes that firm 2 will be active). If product A is very profitable, however, this effect is unlikely to occur.

[10] One special feature of this model, however, is that firm 2 can "concede" to firm 1 only by fully withdrawing from the market. In other models in which concession can be partial, this need not be true. For example, if market competition is of the form described in David M. Kreps and Jose A. Scheinkman (1983) (product production followed by output constrained price competition), then firm 2 will respond to firm 1's more aggressive behavior by reducing its production level, which can make tying profitable even in the absence of complete exclusion.

When tying would drive firm 2 out of the market, firm 1 may or may not find it profitable to do so. The advantage of tying in such an instance is the gain from converting market B from duopoly into a monopoly. The potential loss, however, comes from the fact that firm 1 will be a monopolist who can only offer a bundle. Thus, the presence of a large number of consumers who strongly dislike product $B1$ may make a commitment to bundling unprofitable, even when it leads to exclusion.

At the same time, the welfare consequences of allowing tying in this circumstance are unclear both for consumers and for aggregate efficiency. First, consumers can lose both because of the price effects stemming from the exclusion and also because there is less variety available in market B. The price effect, however, can potentially go either way. The reason is that the same incentive to lower the effective price of good $B1$ that drives firm 2 from the market is also present when firm 1 becomes a monopolist in market B. In general, though, one should expect that if the gains from monopoly in market B are large, the standard price movement should be upward, making consumers uniformly worse off. The effect on aggregate efficiency is still less certain. This is due to two different common welfare ambiguities. First, the biases associated with the free entry process (Spence, 1976; Salop, 1979; Mankiw and Whinston, 1986) imply that exclusion of firms does not necessarily reduce aggregate welfare. Second, it is known from the monopolistic bundling literature (Adams and Yellen, 1976) that bundling in a monopoly setting has ambiguous welfare consequences.

The following example illustrates these points more concretely and also helps to set up the discussion in Section II.

Example 2. Suppose that a consumer of type d has a valuation for good Bi of $v_{Bi} = w - \alpha_i d$, and that d is uniformly distributed on $[0,1]$. Assuming that we have all consumers purchasing from some firm and both firms making sales (so that our earlier assumptions hold in the relevant range), it is straightforward to show that equilibrium prices and profits (gross of fixed costs) in an independent pricing game are given by

$$P_A^0 = \gamma$$

$$P_{Bi}^0 = c_{Bi} + (1/3)$$
$$\times\left[3\alpha_j + (\alpha_i - \alpha_j) + (c_{Bj} - c_{Bi})\right]$$

$$\Pi_1^0 = (\gamma - c_A) + \left[1/9(\alpha_1 + \alpha_2)\right]$$
$$\times\left[3\alpha_2 + (\alpha_1 - \alpha_2) + (c_{B2} - c_{B1})\right]^2$$

$$\Pi_2^0 = \left[1/9(\alpha_1 + \alpha_2)\right]$$
$$\times\left[3\alpha_1 - (\alpha_1 - \alpha_2) - (c_{B2} - c_{B1})\right]^2.$$

In contrast, profits (gross of fixed costs) for firm 2 when firm 1 bundles are given by

$$\Pi_2^0 = \left[1/9(\alpha_1 + \alpha_2)\right]\left[3\alpha_1 - (\alpha_1 - \alpha_2)\right.$$
$$\left. - (c_{B2} - c_{B1}) - (\gamma - c_A)\right]^2,$$

which is lower than in the independent pricing case. Note also that firm 2's profits fall as the surplus associated with good A, $(\gamma - c_A)$, rises. In order to illustrate the other points made above, I consider three special cases of this model in turn: $\alpha_1 = 0$, $\alpha_1 = \alpha_2$, and $\alpha_2 = 0$.

Consider first the case where $\alpha_1 = 0$. In this case, firm 1 always increases its profits by excluding firm 2 (that is, monopoly profits with bundling are greater than duopoly profits with independent good pricing). This is because as a monopolist firm 1 suffers no loss from bundling. Furthermore, the monopoly bundle price of $(w + \gamma)$ leaves all consumers with zero surplus. While all consumers are made worse off, aggregate welfare may rise or fall: if $c_{B2} > c_{B1}$ aggregate welfare must rise since all consumers are still served, and production costs fall. When $c_{B2} < c_{B1}$, the change in aggregate efficiency is given by $\Delta W = K_2 - \alpha_2(c_{B1} - c_{B2})^2$.

For simplicity, assume now that $c_{Bi} \equiv c_B$. When $\alpha_1 = \alpha_2 \equiv \alpha$, the independent pricing equilibrium has full coverage of market B whenever $w > c_B + (3/2)\alpha$. For simplicity, I also assume that $(\alpha/2) < (\gamma - c_A)$, which implies that firm 1 will always sell its bundle to all consumers when it is a monopolist (the qualitative results in the other case are similar). In that case, firm 1's price and profits are given by

$$\bar{P}^0 = w + \gamma - \alpha$$

$$\Pi_1^0 = w + \gamma - \alpha - c_A - c_B.$$

Comparison of these expressions with those for the independent pricing game (setting $\alpha_1 = \alpha_2$) reveals that firm 1 always gains from exclusion. The effective price of good $B1$ $(\bar{P} - \gamma)$, however, falls whenever $(c_B + (3/2)\alpha) \leq w \leq (c_B + 2\alpha)$, so that some consumers (for example, those who were already buying $B1$) are made better off in these cases. Aggregate consumer surplus, however, never rises here: with an independent pricing duopoly, consumer surplus is $w - c_B - (5/4)\alpha$, while it is $(\alpha/2)$ with a bundling monopolist.

Finally, when $\alpha_2 = 0$, firm 1 profits in an independent goods pricing duopoly are given by $\Pi_1^0 = (\gamma - c_A) + (1/3)\alpha_1$ (an interior solution arises whenever $w > c_B + (2/3)\alpha_1$). Then, assuming again that $(\alpha_1/2) < (\gamma - c_A)$, we have that firm 1's profits rise from this exclusion if and only if $w > c_B + (4/3)\alpha_1$. Notice that exclusion is more likely to be profitable as the value of monopolizing market B rises (increases in $(w - c_B)$ and decreases in α_1) and as the competitive constraint that firm 2 imposes when it is in the market becomes more severe (decreases in α_1). ■

II. Heterogeneous Consumer Preferences for Good A

The results of Section I provide two important lessons. First, tying can be profitably used as an exclusionary device. Second, there may be important differences in the likelihood of its use, depending on whether a commitment to tying is possible. A feature of that model, however, was the strong assumption that all consumers have the same valuation of the tying good. In this section, I investigate the effects of relaxing that assumption. Two points emerge from this investigation. First, a commitment to tying need not always result in foreclosure as it did in the model of Section I. Second, when consumer valuations for the tying good differ, tying can be a profitable strategy for firm 1 even in the absence of an ability to commit, and when it is, it may lower firm 2's profitability in a similar manner to that observed earlier. In the following two subsections I consider first the case of commitment and then that of no commitment.

A. *Commitment*

In the model of Section I, firm 1's commitment to offering only a bundle lowered firm 2's sales because it created an incentive for firm 1 to price more aggressively, lowering its bundle price below $P_{B1}^*(P_{B2}) + \gamma$ (strategic foreclosure). More generally, when consumers have heterogeneous preferences for good A, the impact of tying on firm 2's profits can be determined by asking whether, at the bundle price $\bar{P}'$ such that firm 2's sales equal its independent pricing level (i.e., $\bar{P}'$ such that $x^2(\bar{P}' - \gamma, P_{B2}) = x^2(P_{B1}^*(P_{B2}), P_{B2})$), firm 1 has an incentive to lower its price further.[11] This will be true when

$$(3)\quad (\bar{P}' - c_A - c_B)\frac{d\,\text{Bundle Sales}}{d\bar{P}} + x^1(P_{B1}^*(P_{B2}), P_{B2}) < 0.$$

With homogeneous preferences for good A, for example, the inequality in condition (3) is satisfied because $(\bar{P}' - c_A - c_B) > P_{B1}^*(P_{B2}) - c_{B1}$ and $(d\text{ Bundle Sales}/d\bar{P}) = x_1^1(P_{B1}^*(P_{B2}), P_{B2})$.

[11] Here, unlike in Section I, firm 1 may prefer to commit to producing the bundle plus one of the two goods independently as part of an exclusionary strategy (i.e., Lemmas 1 and 2 do not hold here). I focus on the case of a commitment to pure bundling here to provide a comparison with the result in Section I. These other strategies may also lower firm 2's equilibrium profits. If they do so sufficiently to exclude firm 2 from the market, then they will actually be preferred by firm 1 to pure bundling, since they restrict its pricing to a lesser degree when firm 2 is out of the market.

Condition (3) indicates that, with heterogeneous valuations for good A, a commitment to offering only a bundle may fail to lower firm 2's profits for two distinct reasons. First, enough consumers may find good A unattractive (may have valuations below the cost of production) so that firm 1 may have a lower, rather than a higher, margin at price $\bar{P}'$. In such a case, firm 1's monopoly of good A is too weak for bundling to be an effective exclusionary threat in market B; bundling would help rather than hurt firm 2. This effect, of course, is exactly what one should expect a priori.

The second reason is a bit more subtle. As noted above, with homogeneous valuations, the derivative of bundle demand at price $\bar{P}'$ is identical to that arising in market B with independent goods pricing. With heterogeneous valuations, however, this demand derivative can change when firm 1 bundles, potentially counteracting the price-cost margin effect. The clearest example of this occurs in the limiting case where products $B1$ and $B2$ are nearly homogeneous.[12] Then bundling essentially transforms a nearly homogeneous market B into a setting with vertical differentiation (since all consumers value the bundle more than $B2$, but they differ in how large this valuation difference is—see, for example, Avner Shaked and John Sutton, 1982) and can thereby raise firm 2's profits.

The following example, which is an extension of Example 2, illustrates these points.

Example 3. The model considered here is identical to that in Example 2 except that I now allow there to be different possible levels of consumer valuations for product A. I assume that the distribution of γ in the population is described by $F(\gamma)$ and that for all d, $\text{Prob}(\gamma \leq s|d) = F(s)$ (i.e., types are independently distributed across the two markets). In the discussion that follows, I assume that w is large enough so that (in the relevant range) all consumers purchase product B from one of the firms.

Suppose that firm 1 commits to tying by producing only a bundle. A consumer of type (γ, d) will buy the bundle if and only if $d \leq (1/\alpha)[(\bar{P} - \gamma) - P_{B1}]$. It is useful to first assume that for any level of γ some consumers of that type are buying from each of the firms ("interior equilibria"). For interior equilibria, equilibrium profits for the two firms are given by

$$\Pi_1^0 = [1/9(\alpha_1 + \alpha_2)][3\alpha_2 + (\alpha_1 - \alpha_2) + (c_{B2} - c_{B1}) + (E\gamma - c_A)]^2$$

$$\Pi_2^0 = [1/9(\alpha_1 + \alpha_2)][3\alpha_1 - (\alpha_1 - \alpha_2) - (c_{B2} - c_{B1}) - (E\gamma - c_A)]^2,$$

where $E\gamma = \int s dF(s)$. Comparing firm 2's profits to its level under independent goods pricing (derived in Section I), we see that firm 2's profits are lower in the bundling equilibrium as long as $E\gamma > c_A$. Thus, as one would expect, if there are enough consumers who dislike product A (have γ levels below c_A), tying raises rather than lowers firm 2's profits. Relating this finding to condition (3), we see that bundling has no effect on the demand derivative, but that $\bar{P}'$ is now given by $\bar{P}' = P_{B1}^*(P_{B2}) + E\gamma$ so that the inequality in (3) holds if and only if $E\gamma > c_A$.

It is worth noting, however, that the lack of any effect on the demand derivative term in (3) relies heavily on the linearity of the demand structure assumed here. If bundling were to lower this average derivative (in absolute value), this would work against the incentive for more aggressive pricing that arises in this linear model. In fact, this is why interiority of the equilibrium is important for the characterization above. When bundling causes all consumers with some values of γ strictly to prefer either the bundle or good $B2$, this lowers the derivative of firm 1's demand with respect to its bundle price (since none of these consumers are marginal). Intuitively, this effect seems more likely to occur when the dispersion of valuations for good A increases and the

[12] Note that this requires that the K_i's are close to zero if independent pricing would result in a duopoly.

TABLE 1

	$\beta \geq \alpha$	$\beta < \alpha$
$\gamma \geq 3\|\beta - \alpha\|$	$\Pi_2^{\text{BUND}} < \Pi_2^{\text{IND}}$	$\Pi_2^{\text{BUND}} < \Pi_2^{\text{IND}}$
$\gamma \leq 3\|\beta - \alpha\|$	$\Pi_2^{\text{BUND}} < \Pi_2^{\text{IND}}$ if and only if: $\left(\frac{3\beta - \gamma}{3}\right)^2 < \alpha\beta$	$\Pi_2^{\text{BUND}} < \Pi_2^{\text{IND}}$

differentiation between products $B1$ and $B2$ decreases.

To investigate this effect further, consider the special case where γ is uniformly distributed on the interval $(\gamma - \beta, \gamma + \beta)$, where $\beta \leq \gamma$, $c_{B1} = c_{B2} = 0$, and $\alpha_1 = \alpha_2 \equiv \alpha > 0$.[13] In this example, the issue addressed above, that some consumers may value good A at less than its production cost, does not arise (here $E\gamma > c_A$). Rather, the focus here is on the effects of the level of valuation dispersion for good A and the level of product differentiation in market B. Tedious calculations (an example of which is provided in Appendix B) reveal that the effect of a commitment by firm 1 to offering only a bundle on firm 2's profits can be summarized as in Table 1. Examination of the condition in the lower left-hand box (the only case where firm 2's profits are not necessarily lowered by firm 1's bundling) confirms that high levels of dispersion of valuations for good A and low levels of differentiation in market B are necessary for firm 2's profits to rise when firm 1 bundles. Interestingly, though, even when α is close to zero, we need γ not to be too large for this to occur (so that the incentive to make sales of A does not outweigh the differentiation effect).[14] Note also that firm 1's profits may now rise with bundling even if bundling does not drive firm 2 from the market. In fact, in this example, whenever bundling causes firm 2's profits to rise, firm 1's profits rise as well and, further, firm 1's rise in some cases where firm 2's profits fall (this is shown in Appendix B). ■

B. *No Commitment*

The presence of heterogeneous valuations of product A can also cause tying to be firm 1's optimal strategy even in the absence of an ability to commit to this strategy. To see this more clearly, consider first the no commitment game analyzed in Section I. In that game, when both firms are active in market B, firm 1 selects its prices taking firm 2's price as given and acting as a monopolist on the residual demand structure. Given the literature on bundling by multiproduct monopolists (Adams and Yellen, 1976; McAfee, McMillan, and Whinston, 1989), which has found bundling to be a profitable strategy quite generally, it should not be surprising that firm 1 may now find some form of bundling to be its best response to firm 2's price choice. What is interesting from our perspective, however, is that this tying strategy by firm 1 may have detrimental effects on firm 2's profits since, when firm 1 does decide to bundle, it may have an incentive to foreclose sales in market B in a manner similar to that discussed in Section I.

In Whinston (1987), for example, I considered the structure described in Example 3 with two types of valuations for good A, γ_L and γ_H with $\gamma_H > c_A$ and $\text{Prob}(\gamma_H) = \lambda$. For this case I showed that any equilibrium of the no precommitment game is equivalent to an equilibrium of a game where firm 1 is allowed to either sell A and $B1$ independently or to offer only the bundle and

[13] My investigation of this example is motivated in part by the example analyzed in independent work by J. Carbajo, D. DeMeza, and D. J. Seidmann (1987). They illustrate the differentiation effect in an example with homogeneous goods in market B and valuations for goods A and B that are perfectly correlated and uniformly distributed across consumers. Earlier versions of this paper pointed out the implications of noninteriority for the derivative of demand in the context of a two-type (of γ) example.

[14] The reader may be puzzled by this point since it seems that when $\alpha = 0$ firm 2's profits would always rise with bundling. In fact, when $\alpha = 0$, the upper left box would have $\Pi_2^{\text{BUND}} = \Pi_2^{\text{IND}}$ and firm 1 making all sales when it bundles.

product A at price $P_A \in (\gamma_L, \gamma_H]$.[15] In addition, the equilibrium may involve firm 1 pursuing the latter (bundling) strategy, though a necessary condition for this is that $\gamma_L > c_A$ (see Whinston, 1987, for details). When the equilibrium does involve bundling and is "interior" in the sense discussed above, firm 2's profits are

$$\Pi_2^0 = [1/9(\alpha_1+\alpha_2)][3\alpha_1-(\alpha_1-\alpha_2) \\ -(c_{B2}-c_{B1})-(1-\lambda)(\gamma_L-c_A)]^2.$$

Thus, when firm 1 does tie here, it forecloses firm 2's sales in a similar manner to that observed earlier. Note, though, that firm 2's equilibrium profits are larger here than when firm 1 commits to only offering a bundle. The reason is that when firm 1 also offers product A independently, it is assured of making sales of product A to all type H consumers regardless of whether they buy product $B1$; thus, here the incentive for foreclosure arises only from the L types and firm 2's profits fall only if $\gamma_L > c_A$.

Though the effect of firm 1's tying here may be exclusionary (firm 2, anticipating that firm 1 will tie, may choose to be inactive), one might argue that its motives are in some sense "innocent" since its decision to tie is never affected by the possibility that firm 2 might be excluded from the market. Such dynamic considerations, however, may be important even when firm 1 cannot precommit to tying. For example, if firm 2 faces a financial constraint that it must meet in order to remain active in the market (as in the work of J. P. Benoit, 1984; and Drew Fudenberg and Jean Tirole, 1986), firm 1 may be led to use tying in order to lower firm 2's profits and increase the likelihood that firm 2 will be forced to exit the market, even when tying is not profit-maximizing in a static sense.

To formalize this idea, consider a simple extension of the earlier no commitment model in which there are two production periods. If firm 2 decides to be active and incurs the set-up cost K_2 prior to period 1, it may face a financial constraint that it must meet after period 1 in order to be able to remain in the market in period 2. In particular, suppose that with probability $1-\theta$ firm 2 will not face a financial constraint, while with probability $\theta f(\Pi)$ firm 2 will face a constraint that prohibits continued participation if first-period profits were less than Π and assume that $f'(\Pi) \geq 0$ (there is a diminishing marginal return to predation).

In this setting, what is the effect of an increase in θ on the attractiveness of tying for firm 1? It is not difficult to see that for the two type example if $\gamma_L > c_A$ (and outcomes are "interior") then increases in θ make tying a relatively more attractive policy for firm 1 in period 1 for any given level of P_{B2}. The central (and very general) idea is that increases in θ make firm 1 care more about foreclosure relative to current profits.

To see this more formally, let G denote the benefit to firm 1 if firm 2 does not meet its financial constraint and fix some initial level of θ and P_{B2}. Suppose, first, that firm 1 pursues its best independent pricing policy and that this results in a profit level for firm 2 of Π_2^I. Then, firm 1's price choices are equal to the level that it would choose in the simple one production period model if its marginal costs of production for $B1$ were $c_{B1}-\theta Gf(\Pi_2^I)(P_{B2}-c_{B2})$ instead of c_{B1}. Likewise, if firm 1 pursues its optimal bundling strategy and thereby gives firm 2 profits of Π_2^B, then its prices are equal to those it would pick in the static game if its marginal cost was

$$c_{B1}-\theta Gf(\Pi_2^B)(P_{B2}-c_{B2}).$$

Since we have seen that the optimal bundling strategy in the one period no pre-

[15] That is, we can without loss of generality restrict firm 1's pricing strategy choices to one of these two forms. This equivalence actually holds for any market B structure that satisfies the assumptions made in Section II (a proof of this fact is available from the author upon request). It is worth noting that a bundling strategy of this sort may not appear to be tying at all since firm 1 does offer to sell product A at a price that some consumers are willing to pay. For type L consumers, however, this offer is unattractive, putting them in exactly the same situation as when firm 1 offers only a bundle.

commitment game results in lower profits for firm 2 then does the optimal independent pricing policy for any given level of c_{B1} (since $\gamma_L > c_A$), it must be that $\Pi_2^B < \Pi_2^I$ (that is, that bundling leads to foreclosure). But the envelope theorem then implies that a small increase in θ raises the profits from the optimal bundling best response by more than it raises the profits from the optimal independent pricing best response (since the derivative of firm 1 profits with respect to θ is $GF(\Pi_2)$). Thus, in this example, increases in θ strictly increase the likelihood that firm 1 will find bundling to be its best response (since bundling is never optimal if $\gamma_L < c_A$).[16]

III. Complementary Products

I now turn to the case of complementary products used in fixed proportions. I first consider a model of fixed proportions that is essentially an extension of the simple example quoted above from Posner (1976) to the case where the tied good market involves differentiated products with scale economies in production and an oligopolistic, rather than a competitive, market structure. Despite these differences, I show that Posner's central contention continues to hold: a monopolist of one component never finds it worthwhile to tie in order to reduce the level of competition in the market for the other component. The key point is that with complementary products used in fixed proportions, the monopolist can actually derive *greater* profits when its rival is in the market than when it is not because it can benefit through sales of its monopolized product from the additional surplus that its rival's presence generates (due to product differentiation).

Nevertheless, I then show that in two natural extensions of this model in which the monopolized product is no longer essential for all uses of other components, tying once again emerges as a profitable exclusionary strategy. In one case, the presence of an inferior, competitively supplied alternative to the "monopolized" component leads to results that parallel those of the independent products case. In the other case, the existence of a second use for the nonmonopolized product (such as a replacement part market) can give the monopolist an incentive to tie in order to eliminate competition in this other market.

The discussion in the text focuses on the case of precommitment. In fact, for each of the models considered here, any no precommitment outcome is equivalent to an equilibrium of the independent pricing game.[17] Of course, this is therefore also true when firm 1 produces A and $B1$ independently in the commitment game. In order to simplify the exposition, in Parts B and C below, I will use this fact and simply compare bundling outcomes to the independent pricing game equilibria when investigating whether firm 1 would find a commitment to bundling to be a profitable exclusionary device.

A. *The Basic Model*

Consider the following simple model. There are two components needed to comprise a system, A and B: a system consists of one unit of each. As before, firm 1 is a monopolist of component A, and two different versions of component B could potentially be available, $B1$ and $B2$. The production technology for these products is as before.

[16] The two-type of γ example considered here is special in one sense. With more general distributions of γ, firm 1's best response in the one period no precommitment game will quite generally involve some form of bundling (see McAfee, McMillan, and Whinston, 1989). In such cases, one would have to examine how increases in θ affected the degree of bundling (i.e., the difference between $\bar{P}$ and $P_A + P_{B1}$).

[17] The proofs of this fact for the three models presented in this section are available from the author upon request. For the model of Part C, the result requires the use of Selten's (1975) notion of trembling-hand perfection in order to eliminate the use of weakly dominated strategies. In Parts B and C this equivalence is a consequence of the homogeneity of valuations assumed there (as in Section I).

The set of consumers is the same as in Section I. Each consumer demands at most one unit of the system. A consumer of type d's valuation of a system with product Bi is $v_{A/Bi}(d)$. When goods A, $B1$, and $B2$ are independently priced, consumers' demand for an A/Bi system is given by some function $x^i(P_A + P_{B1}, P_A + P_{B2})$, where $x^i_j(\cdot,\cdot) \geq 0$ if $i \neq j$ and ≤ 0 if $i = j$, with strict inequalities whenever $x^i(\cdot,\cdot) \in (0,1)$, and where $(x^1_i(\cdot,\cdot) + x^2_i(\cdot,\cdot)) \leq 0$.

In the case of independent products we implicitly assumed that purchase of a produced bundled unit allowed the independent use of either of the products (the proof of Lemma 1, for example, uses this fact). Though natural in the case of independent products, this assumption is less so when products must be used together. For example, the bundling of a stereo tuner and a stereo amplifier into a stereo receiver may not allow the buyer to use just the amplifier in conjunction with another manufacturer's tuner. Thus, here I assume that production of a bundled good does not allow the user to use only part of the bundle.

In this model, since component A is essential to any system, firm 1 is trivially able to exclude firm 2 by committing to produce only a bundle. Nevertheless, as the following proposition indicates, firm 1 never finds it worthwhile to tie in order to exclude firm 2.

PROPOSITION 3: *If a commitment to tying causes firm 2 to be inactive, firm 1 can do no worse—and possibly better—by committing to producing only independent components.*

PROOF:

Suppose that firm 1's precommitment to tying (by not producing one or both of the components individually) causes firm 2 to be inactive. In this case, since only firm 1's bundle price is relevant once firm 2 is inactive, firm 1's profits given its optimal bundle price of $\bar{P}^*$ are

$$(\bar{P}^* - c_A - c_{B1})x^1(\bar{P}^*, \infty).$$

Suppose that firm 1 instead commits to only producing components A and $B1$ independently. One pricing policy that it can always follow, regardless of whether firm 2 is active, is to set individual component prices of $\hat{P}_{B1} = c_{B1} - \varepsilon$ (where $\varepsilon > 0$) and $\hat{P}_A = \bar{P}^* - \hat{P}_{B1}$. If firm 2 is inactive, this pricing scheme leads to exactly the same level of profits as did the bundling outcome. If firm 2 is active, however, firm 1's profits will be at least as large as those in the bundling outcome since they are given by

$$(\bar{P}^* - c_A - c_{B1})\left[x^1(\bar{P}^*, \hat{P}_A + P_{B2}) + x^2(\bar{P}^*, \hat{P}_A + P_{B2})\right] + \varepsilon x^2(\bar{P}^*, \hat{P}_A + P_{B2}),$$

when firm 2 names price P_{B2} (since $x^1(\cdot,\cdot) + x^2(\cdot,\cdot)$ weakly increases when prices fall and $x^2(\cdot,\cdot)$ is nonnegative). □

The basic idea behind this result is fairly simple to see. If firm 2 did not exist, firm 1 could do as well as it does through bundling by setting independent prices that had component $B1$ priced at or below cost and component A's price set at a high level; it would simply earn all of its profits on sales of component A (consumers' purchases depend only on the sum of the prices). But, if pricing in this manner leads firm 2 to be active, this can only raise firm 1's profits since firm 1 would then sell more component A's (on which it makes profits) and fewer component B's (on which it has a negative margin). Intuitively, firm 1 is able to benefit through sales of its product A from the increase in surplus generated by firm 2's presence.

While firm 1 never gains from committing to tying here if this forces firm 2 to be inactive, firm 1 may commit to tying in order to price discriminate. For example, suppose that some set of consumers get positive benefits only out of an $A/B1$ system, while the remainder get positive benefits only out of an $A/B2$ system, and that the latter group's valuation of its desired system is much higher. Then firm 1 will want to set a very high price for good A in order to extract surplus from this latter group, and a very low price for good $B1$ in order to get an optimal $A/B1$ system price for the former group. If this attempt hits the nonnegativity constraint on P_{B1}, however, then firm 1 will find it worthwhile to tie by offering a bundle with price $\bar{P} < P_A$.[18]

Firm 1's lack of desire to use tying as an exclusionary device can change dramatically, however, when firm 1's monopolized component is not essential for all uses of product $B2$. I now consider two natural extensions of the above model in which tying can prove to be not only an effective exclusionary device but also a profitable one.

B. *An Inferior, Competitively Supplied Component A: Strategic Foreclosure*

Suppose that there exists a uniformly inferior, competitively supplied alternative to firm 1's product A, denoted as product $A2$ (henceforth, firm 1's product A will be denoted by $A1$). The cost of component $A2$ is also c_A, but compared with the valuations described above for $A1/B1$ and $A1/B2$ systems, a consumer's valuation for a system that has product $A2$ in it rather than $A1$ is $(\gamma - c_A)$ lower (i.e., $v_{A2/Bi}(d) = v_{A1/Bi}(d) - (\gamma - c_A)$) where $\gamma > c_A$.

Consider, first, the independent pricing game (which, as noted above, yields an outcome identical to what occurs if firm 1 produces A and $B1$ only independently). In this game, firm 1 always sets $P_{A1} \leq \gamma$ and makes all component A sales. When firm 1 sets $P_{A1} < \gamma$ in this equilibrium, the inferior alternative (product $A2$) is irrelevant for pricing and profits. In the case where $P_{A1} = \gamma$, however, the presence of the inferior product $A2$ constrains firm 1's equilibrium pricing and profits. This could mean that, contrary to Proposition 3, firm 1 would prefer to have firm 2 out of the market (firm 1 can no longer necessarily benefit through its component $A1$ sales from the surplus created by the presence of firm 2).[19] Example 4 illustrates this point and shows how the presence of component $A2$ can make competitive interaction here look very much like the independent products case considered earlier.

Example 4. Suppose that $v_{A1/Bi}(d) = w - \alpha d$, $c_A > 0$ and $c_{B1} = c_{B2} \equiv c_B > 0$, and that $w \geq 2\alpha + c_A + c_B$ (to ensure that all consumers buy a system; note the parallel to Example 2). Ignoring the constraint imposed by the presence of product $A2$, the independent pricing equilibrium level of P_{A1} is increasing in w.[20] When $w > \gamma + c_B + (3/2)\alpha$, the unique equilibrium involves prices of $P_A^0 = \gamma$ and $P_{B1}^0 = P_{B2}^0 = c_B + \alpha$, and all consumers receive positive surplus (see Figure 2). Profits (gross of fixed costs) are given by

$$\Pi_1^0 = (\gamma - c_A) + (\alpha/2)$$

$$\Pi_2^0 = (\alpha/2).$$

Note that this equilibrium essentially replicates the independent goods outcome from Section II (Example 2 with $\alpha_1 = \alpha_2$ and $c_{B1} = c_{B2}$). That is, the presence of a competitive constraint from product $A2$ serves to "uncouple" the two component markets. As in the independent products case, if w is large, firm 1's profits are increased by firm 2 being inactive (firm 1 then acts as a systems

[18] This point is analogous to the observation that an upstream monopolist may wish to integrate vertically forward into one of the industries that uses its product in order to achieve price discrimination across users (see, for example, Tirole, 1988, p. 141). Note that bundled production is essential for this purpose since otherwise the second set of consumers would buy the bundle to get their component A whenever the bundle price was lower than the price of good A alone.

[19] The fact that the presence of an inferior competitively supplied product A can potentially prevent firm A from deriving maximal (two-product monopoly) profits has also been noted in Ordover, Sykes, and Willig (1985).

[20] More precisely, though multiple equilibria exist in the game when component $A2$ does not exist (corresponding to a range of values for P_{B1} and P_{B2} that is independent of w), in any such equilibrium the level of P_A is given by $P_A^* = w - (1/2)[\alpha + c_B - 3P_{B1}^*]$.

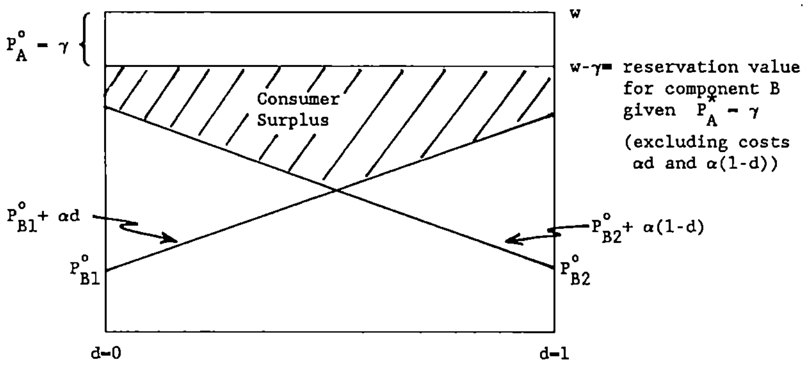

P°A − γ
w
w-γ= reservation value for component B given P*A = γ (excluding costs αd and α(1-d))
Consumer Surplus
P°B1+ αd
P°B1
P°B2+ α(1-d)
P°B2
d=0
d=1

FIGURE 2

monopolist setting an $A1/B1$ system price of w).[21] ■

Now consider the commitment game. When firm 1 would prefer firm 2 to be out of the market, can a commitment to tying by firm 1 force firm 2 out of the market (note that firm 1's component $A1$ is not essential)? The answer is yes, and for the same basic reason as in Section I: when it is only able to sell a bundle, firm 1 can only gain its profits from component $A1$ if it also sells component $B1$; this causes firm 1 to foreclose sales in the component B market. To see this formally, suppose that firm 1 can only produce a bundle. When the presence of product $A2$ constrains firm 1's pricing in an independent pricing game (so that $P_{A1}=\gamma$), firm 1's price for component $B1$ given firm 2's price P_{B2} is given by $P_{B1}^*(P_{B2})$ such that

$$(4)\quad \{[P_{B1}^*(P_{B2})-c_{B1}]x_1^1(\gamma+P_{B1}^*,\gamma+P_{B2}) + x^1(\gamma+P_{B1}^*,\gamma+P_{B2})\} + (\gamma-c_A)[x_1^1(\gamma+P_{B1}^*,\gamma+P_{B2}) + x_1^2(\gamma+P_{B1}^*,\gamma+P_{B2})]=0.$$

The first term of this expression represents the effect on sales of component $B1$ of marginally changing P_{B1}, while the second is the effect on sales of component $A1$. Note that this second change is due to the *total* change in system sales. In contrast, when firm 1 commits to bundling, its optimal bundle price given firm 2's price, P_{B2}, is given by $\bar{P}^*(P_{B2})$ such that

$$(5)\quad [\bar{P}^*(P_{B2})-c_A-c_{B1}]x_1^1(\bar{P}^*,\gamma+P_{B2}) + x^1(\bar{P}^*,\gamma+P_{B2})=0.$$

At $\bar{P}=P_{B1}^*(P_{B2})+\gamma$, this expression becomes

$$(6)\quad \{[P_{B1}^*(P_{B2})-c_{B1}]x_1^1(\gamma+P_{B1}^*,\gamma+P_{B2}) + x^1(\gamma+P_{B1}^*,\gamma+P_{B2})\} + (\gamma-c_A)x_1^1(\gamma+P_{B1}^*,\gamma+P_{B2}).$$

Note that $x_1^2(\cdot,\cdot)$ does not appear in the second term of (6). This represents the fact that when firm 1 bundles, only by increasing sales of $A1/B1$ systems does it increase the sales of component $A1$. Firm 1 is therefore led to set $\bar{P}^*(P_{B2})<P_{B1}^*(P_{B2})+\gamma$, foreclosing sales in market B and lowering firm 2's profits.[22] Thus, by committing firm 1 to

[21] Unlike the independent products case, firm 1 suffers no loss from bing restricted to bundle when it is a monopolist. Rather, here the cost of exclusion of firm 2 is that firm 1 is unable to capture any of the surplus created by firm 2 (through firm 1's sales of component A).

[22] This incentive for foreclosure is similar to the effects studied in Carmen Matutes and Pierre Regibeau (1986). They study product compatibility in a symmetric duopoly and identify a collusive incentive to

"strategic foreclosure," tying can exclude firm 2 from market B and thereby raise firm 1's profits.[23,24]

Example 4 *cont*. If firm 1 commits to only producing a bundle, firm 2's equilibrium profit when both firms are active is given by

$$\Pi_2^0 = \max\left\{0,(1/2\alpha)\left(\alpha - \frac{\gamma - c_A}{3}\right)^2\right\},$$

which is lower than its profit under independent pricing. Note that if firm 1 bundles and forces firm 2 to be inactive, all consumers receive zero surplus here (although, as usual, aggregate welfare may either fall or rise). ■

C. *An Alternative Use for Product B: Direct Foreclosure*

Next, consider an alternative variation in the basic model. Suppose that there exists an alternative use for component B that does not rely on the simultaneous purchase of component A. One example of such a use is a replacement parts market for existing owners of a system who need to replace only component B. Because component A is not essential for the use of product B in that market, firm 1 is not able to benefit from firm 2's presence in this market through sales of good A and the logic of Proposition 3 therefore breaks down. Firm 1 may now find it worthwhile to exclude firm 2, if it can, in order to monopolize this other market for product B. Furthermore, because component A is still essential for certain uses of product B, firm 1 may have the means to accomplish this end: by offering to sell compoent A only in a bundle with component $B1$, firm 1 directly forecloses firm 2's sales in the joint use market (foreclosure of these sales is complete regardless of firm 1's bundle price), which may drive firm 2's profits below the level that justifies its continued operation. The following simple example illustrates these points.

Example 5. Suppose that there are two types of consumers. Type I consumers desire a system. There are a continuum of type I consumers indexed by the uniformly distributed variable $d \in [0,1]$ with total measure 1. Consumer d has valuations for the two possible systems of $v_{A/B1}(d) = w \cdot d$ and $v_{A/B2}(d) = w \cdot d + \gamma_1$. Type II consumers, of which there are a total measure of θ, only desire product B. Each type II consumer has valuations for products $B1$ and $B2$ of $v_{B1} = \varphi$ and $v_{B2} = \varphi + \gamma_2$. The firms are unable to discriminate (in a third degree sense) across these consumers in their pricing. The cost structure has $c_A > 0$, $c_{B1} = c_{B2} \equiv c_B > 0$, $K_2 > 0$, and $K_1 = 0$. Finally, I make two further assumptions:

$$\text{(A1)} \qquad (1+\theta)\gamma_2 > \gamma_1 > \gamma_2$$

and

$$\text{(A2)} \qquad w > \max\{4\gamma_2 - \gamma_1 + c_A + c_B, \; \gamma_1 + c_A + c_B\}.$$

have compatibility. This corresponds to the "puppy dog" strategy in the Fudenberg and Tirole (1984) taxonomy for the case of accommodation under price competition, in contrast to the "top dog" strategy that I focus on here (see fn. 7).

[23] The discussion in the text has only compared producing A and $B1$ independently with producing only a bundle. One might wonder about other alternatives. In fact, it can be shown that as long as the sort of price discrimination motivation discussed in Part A is not present, any of the other alternatives are either equivalent to independent pricing (Bundle and $A1$; Bundle, $A1$, and $B1$), equivalent to producing only a bundle (Bundle and $B1$), or clearly inferior to these options ($A1$ only, $B1$ only).

[24] A long-standing issue in the legal treatment of tying is when to treat the tying and tied products as distinct products. A common argument is that the tied product must be one that consumers might want to purchase separately, without also purchasing the tying product. In Justice O'Connor's concurrence in *Jefferson Parish Hospital District No. 2 v. Hyde*, 466 U.S. 2 (1984), for example, she argues this position because "When the tied product has no use other than in conjunction with the tying product, a seller of the tying product can acquire no *additional* market power by selling the two products together." The model analyzed here illustrates that this view is incorrect unless one defines "other uses," contrary to Justice O'Connor's meaning, to include use with other producers' component A's.

Consider, first, the outcome of the independent pricing game. The unique equilibrium outcome when both firms are active involves prices of[25]

$$P_{B1}^0 = c_B$$

$$P_{B2}^0 = c_B + \gamma_2$$

$$P_A^0 = [w + (\gamma_1 - \gamma_2) + c_A - c_B]/2.$$

In this equilibrium, all consumers buying a component B buy product $B2$, and profits for the two firms are given by

$$\Pi_1^0 = [w + (\gamma_1 - \gamma_2) - c_A - c_B]^2/4w$$

$$\Pi_2^0 = \gamma_2 \cdot \left[\theta + \frac{w + (\gamma_1 - \gamma_2) - c_A - c_B}{2w}\right] - K_2.$$

Suppose, instead, that firm 1 commits to producing only a bundle and product $B1$ alone. In this case the unique equilibrium prices when firm 2 is active are given by

$$P_{B1}^0 = c_B$$

$$P_{B2}^0 = c_B + \gamma_2$$

$$\bar{P}^0 = (w + c_A + c_B)/2,$$

and profits are

$$\Pi_1^0 = (w - c_A - c_B)^2/4w$$

$$\Pi_2^0 = \gamma_2 \cdot \theta - K_2.$$

Thus, by committing to tie, firm 1 denies firm 2 its profitable sales to type I consumers, lowering firm 2's profits, and possibly forcing firm 2 to be inactive. Furthermore, if tying does force firm 2 to be inactive, firm 1's profit is $((w - c_A - c_B)^2/4w) + \theta(\varphi - c_B)$, which is larger than its independent pricing profits if φ, the gain from monopolizing the type II market, is large.[26] Finally, if firm 1 does exclude firm 2 in this manner, all consumers are made worse off here, although aggregate welfare may either fall or rise. ■

IV. Conclusion

The above results demonstrate, in my view, that the leverage hypothesis can be formally modeled in a coherent and appealing way. Once one allows for scale economies and strategic interaction, tying can make continued operation by a monopolist's tied market rival unprofitable by leading to the foreclosure of tied good sales. As the models above have indicated, such a strategy can be a profitable one for a monopolist, often precisely because of this exclusionary effect on market structure.

While the analysis vindicates the leverage hypothesis on a positive level, its normative implications are less clear. Even in the simple models considered here, which ignore a number of other possible motivations for the practice, the impact of this exclusion on

[25] I am ignoring equilibria here that involve firm 1 pricing its component $B1$ below cost and making no sales. As earlier, these equilibria involve the use of a weakly dominated strategy by firm 1 and can be eliminated through the use of Selten's (1975) notion of trembling-hand perfection.

[26] The reader may be wondering about other alternatives available to firm 1. A commitment to producing A only, $B1$ only, or just a bundle is worse as an exclusionary strategy for firm 1 than committing to produce the bundle and $B1$ since firm 2's profits are higher when it is active under these strategies than when firm 1 commits to produce the bundle and $B1$, and firm 1's profits are lower under these options if firm 2 is inactive. They also are less attractive as an accommodation strategy for firm 1 than independent production of A and $B1$. Producing A, $B1$, and a bundle yields an outcome equivalent to the independent production outcome (restricting attention to trembling-hand perfect equilibria). Finally, producing a bundle and A is less effective as an exclusionary strategy than producing a bundle and $B1$ (it gives firm 2 higher profits if it is active and firm 1 lower profits when firm 2 is not active), and when firm 2 is active, no pure strategy (trembling-hand perfect) equilibrium with this product offering can give firm 1 higher profits than when it produces A and $B1$ independently. However, a pure strategy equilibrium may not exist here. A sufficient condition for a pure strategy equilibrium to exist is that $\gamma_2 < (\varphi - c_B)$. Thus, when this condition holds, firm 1 can effectively limit itself to the two options considered in the text.

welfare is uncertain. This fact, combined with the difficulty of sorting out the leverage-based instances of tying from other cases, makes the specification of a practical legal standard extremely difficult.

Finally, it should be noted that the leverage debate is not limited to the practice of tying, but rather arises in numerous areas of antitrust analysis. With the practice of reciprocity, for example, a monopsonistic buyer of some product refuses to buy from his suppliers unless they also buy a product (in which he may face competition) from him. Alternatively, when a vertically integrated monopolistic input supplier can sell his input to both his own downstream manufacturer and to a rival manufacturer, a refusal to supply this rival manufacturer is similar to the tying of complementary goods. The results hère raise the possibililty that the use of leverage as an effective and profitable exclusionary device could arise in these other settings as well.[27]

APPENDIX A

PROOF OF LEMMA 2: Suppose, first, that firm 2 is active and that the equilibrium prices are $((P^0_{B1}, \bar{P}^0); P^0_{B2})$. There are two cases to consider. If $\bar{P}^0 \leq P^0_{B1} + \gamma$, then all consumers prefer the bundle to buying only good $B1$ from firm 1 (again, for expositional reasons, I assume here that consumers buy the bundle when they are indifferent). If so, then firm 1 selling only the bundle at price $\bar{P}^0$ generates identical sales and profits for both firms for all P_{B2}. If $\bar{P}^0 \geq P^0_{B1} + \gamma$, then it must be that firm 1 is making no sales since otherwise it could do better by setting $\bar{P} = P^0_{B1} + \gamma$ (it would make exactly the same number of sales of the bundle as it did of $B1$, but at a larger margin since $\gamma > c_A$). In this case, firm 1 selling only the bundle at price $P^0_{B1} + \gamma$ generates identical sales and profits for both firms when $P_{B2} = P^0_{B2}$ and for firm 2 for all P_{B2}. A similar argument holds if firm 2 is not active. Therefore, any perfect equilibrium outcome (including decisions regarding activity in market B) is equivalent to one that arises after firm 1 has committed to producing only the bundle.

PROOF OF PROPOSITION 2: The argument is a simple comparative statics exercise. Letting $\phi \equiv \bar{P}_1 - \gamma$, firm 1's problem, given P_{B2}, can be written

$$\max_{\phi} [(\phi - c_{B1}) + (\gamma - c_A)] \cdot x^1(\phi, P_{B2}).$$

The bundling equilibrium is then characterized by the following two equations, which have a unique solution (with positive sales by both firms) under our assumptions.

$$[(\phi^{**} - c_{B1}) + (\gamma - c_A)] x^1_1(\phi^{**}, P^{**}_{B2}) + x^1(\phi^{**}, P^{**}_{B2}) = 0$$

$$(P^{**}_{B2} - c_{B2}) x^2_2(\phi^{**}, P^{**}_{B2}) + x^2(\phi^{**}, P^{**}_{B2}) = 0.$$

Note that if $\gamma = c_A$, then $(\phi^{**}, P^{**}_{B2}) = (P^*_{B1}, P^*_{B2})$, the independent pricing equilibrium. Now define (omitting arguments of functions):

$$A \equiv 2x^1_1 + [(\phi^{**} - c_{B1}) + (\gamma - c_A)] x^1_{11}$$

$$B \equiv 2x^2_2 + (P^{**}_{B2} - c_{B2}) x^2_{22}$$

$$C \equiv x^1_2 + x^1_{12}[(\phi^{**} - c_{B1}) + (\gamma - c_A)]$$

$$D \equiv x^2_1 + x^2_{21}(P^{**}_{B2} - c_{B2}).$$

The assumption that $P^{*\prime}_{Bi}(P_{Bj}) \in (0,1)$ implies that $(A, B) \ll (-C, -D) \ll 0$. This then implies that

$$\text{sign} \frac{d\phi^{**}}{d\gamma} = \text{sign}\{-Bx^1_1\} < 0$$

$$\text{sign} \frac{dP^{**}_{B2}}{d\gamma} = \text{sign}\{Dx^1_1\} < 0,$$

so that both firms' profits fall relative to the independent pricing equilibrium. □

APPENDIX B

Here I work out the example for the cases where $\beta \geq \alpha$. In this class of cases, the division of consumers between the two firms can be represented diagrammatically as in Figure 4. Consider first equilibria that are in region (i), that is, that satisfy $\bar{P} \leq (\gamma - \beta) + (P_{B2} + \alpha)$. In this region, the first-order conditions for the two firms are as follows (these conditions are sufficient for a maximum since at any point where these conditions

[27] In fact, the models analyzed above can frequently be reinterpreted to apply to these other settings. Some differences do arise, however, in modeling the various practices. For example, the extent to which commitment is possible is likely to vary by practice. Also, in the case of vertical integration discussed in the text, the upstream monopolist sells his component (input) to the unintegrated final goods producer, who then sets a price for the entire system (finished good) rather than selling directly to consumers who put a system together themselves. Furthermore, in such a setting, more complicated wholesale contracts may be possible than the simple linear pricing considered here.

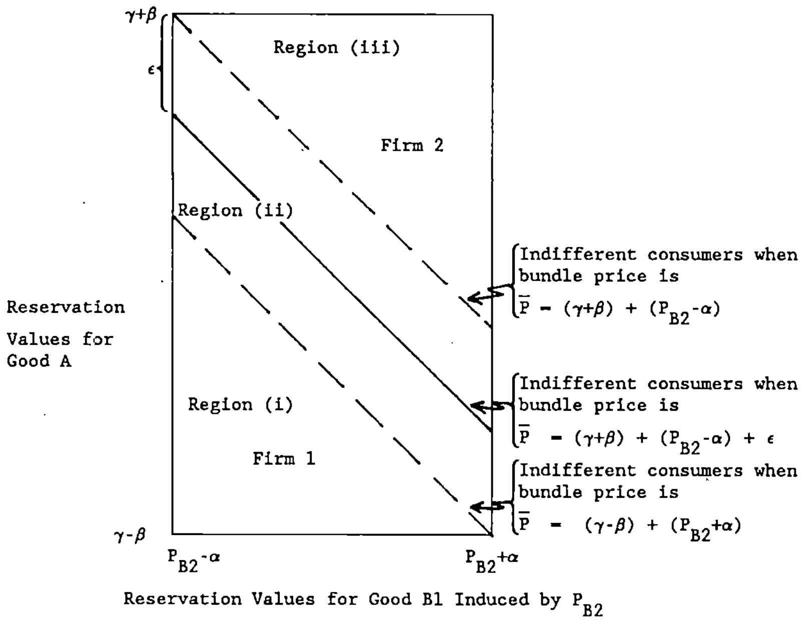

γ+β
ε
γ-β
Region (iii)
Firm 2
Region (ii)
Region (i)
Firm 1
Reservation Values for Good A
Indifferent consumers when bundle price is P̄ - (γ+β) + (P_B2-α)
Indifferent consumers when bundle price is P̄ - (γ+β) + (P_B2-α) + ε
Indifferent consumers when bundle price is P̄ - (γ-β) + (P_B2+α)
P_B2-α
P_B2+α
Reservation Values for Good B1 Induced by P_B2

FIGURE 3

hold, the firms' profit functions are concave):

$$\text{Firm 1:}\quad 4\alpha\beta-(1/2)[\bar{P}-(\gamma-\beta)-(P_{B2}-\alpha)]^2$$
$$-\bar{P}[\bar{P}-(\gamma-\beta)-(P_{B2}-\alpha)]=0$$

$$\text{Firm 2:}\quad [\bar{P}-(\gamma-\beta)-(P_{B2}-\alpha)]-2P_{B2}=0.$$

From firm 2's first-order condition we see that we are in region (i) if and only if $P_{B2}\le\alpha$. Solving the two first-order conditions for P_{B2} yields the following expression:

$$-8(P_{B2})^2+2[\alpha-(\gamma-\beta)]P_{B2}+4\alpha\beta=0.$$

This expression is strictly concave and is nonnegative at $P_{B2}=0$. Hence, $P_{B2}\le\alpha$ if and only if the value of this expression is nonpositive at $P_{B2}=\alpha$. Substituting yields the requirement that $\gamma\ge 3(\beta-\alpha)$. Firm 2's profits in this region under bundling are $(1/2\alpha\beta)(P_{B2})^3$ compared with its profits of $(\alpha/2)$ under independent goods pricing. Since $P_{B2}\le\alpha$ in this region, firm 2's profits must fall.

Consider now bundling equilibria that fall in region (ii), that is, where $\bar{P}\in((\gamma-\beta)+(P_{B2}+\alpha),\ (\gamma+\beta)+(P_{B2}-\alpha))$. Straightforward analysis of the firms' first-order conditions reveals that in equilibrium we must have $3P_{B2}=3\beta-\gamma$. In addition, to be in region (ii), P_{B2} must satisfy $2\beta-\alpha\ge P_{B2}\ge\alpha$, or substituting for P_{B2}: $3(\beta-\alpha)\ge\gamma\ge 3(\alpha-\beta)$. The first of these inequalities is just the reverse of our region (i) condition, while the second, which assures that we are not in region (iii), is always satisfied since $\beta\ge\alpha$ (in fact, the bundling equilibrium can never be in region (iii)). Firm 2's profits under bundling in this region are given by $(1/2\beta)(P_{B2})^2$ compared with $(\alpha/2)$ under independent goods pricing. Substituting for P_{B2} yields the condition in the text. Firm 1's profits under bundling in this region are given by

$$\Pi_1=\bar{P}\{1-(1/2\beta)[\bar{P}-(\gamma-\beta)-P_{B2}]\}$$
$$=\left(\frac{3\beta+\gamma}{3}\right)\left(\frac{1}{2}+\frac{\gamma}{6\beta}\right),$$

while under the parameter values of this region its independent goods pricing profits are given by

$$(\alpha/2)+\left(\frac{\gamma+\beta}{2}\right)^2(1/2\beta).$$

Bundling then yields firm 1 larger profits than independent pricing (assuming that firm 2 remains active) if and only if

$$\left(\frac{9\beta+5\gamma}{6}\right)\left(\frac{3\beta-\gamma}{6}\right)\ge\alpha\beta.$$

But the expression on the left side of this inequality is strictly larger than $((3\beta-\gamma)/3)^2$, which implies that whenever firm 2's profits are higher under bundling, so are firm 1's. There is also clearly an area of the parameter space where firm 1 is better off and firm 2 is worse off under bundling compared to independent goods pricing.

The analysis of cases where $\alpha > \beta$ proceeds in a similar manner.

REFERENCES

Adams, W. J. and Yellen, J. L., "Commodity Bundling and the Burden of Monopoly," *Quarterly Journal of Economics*, 1976, *90*, 475–98.

Benoit, J. P., "Financially Constrained Entry in a Game with Incomplete Information," *Rand Journal of Economics*, Winter 1984, *15*, 490–99.

Blair, Roger D. and Kaserman, David L., "Vertical Integration, Tying, and Antitrust Policy," *American Economic Review*, June 1978, *68*, 397–402.

______ and ______, *Antitrust Economics*, Homewood, IL: Richard D. Irwin, 1985.

Bork, Robert H., *The Antitrust Paradox*, New York: Basic Books, 1978.

Bowman, W. S., "Tying Arrangements and the Leverage Problem," *Yale Law Review*, November 1957, *67*, 19–36.

Bulow, Jeremy I., Geanakoplos, John D. and Klemperer, Paul D., "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy*, June 1985, *93*, 488–511.

Burstein, Meyer L., "The Economics of Tie-in Sales," *Review of Economics and Statistics*, February 1960, *42*, 68–73.

Carbajo, J., DeMeza, D. and Seidmann, D. J., "A Strategic Motivation for Commodity Bundling," London School of Economics, mimeo., 1987.

Director, Aaron and Levi, Edward, "Law and the Future: Trade Regulation," *Northwestern University Law Review*, 1956, *51*, 281–96.

Fisher, Franklin M., McGowan, John J. and Greenwood, Joen E., *Folded, Spindled, Mutilated: Economic Analysis and U.S. v. IBM*, Cambridge, MA: MIT Press, 1983.

Fudenberg, Drew and Tirole, Jean, "The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look," *American Economic Review*, May 1984, *74*, 361–66.

______ and ______, "A Signal-Jamming Theory of Predation," *Rand Journal of Economics*, Autumn 1986, *17*, 366–77.

Gelman, J. R., "Tie-ins Involving Bundles with Fixed Proportions Demand," Federal Trade Commission Working Paper No. 97, 1983.

Kaplow, Louis, "Extension of Monopoly Power Through Leverage," *Columbia Law Review*, April 1985, *85*, 515–54.

Krattenmaker, T. G. and Salop, S., "Anticompetitive Exclusion: Raising Rivals' Costs to Achieve Power Over Price," *Yale Law Journal*, December 1986, *96*, 209–93.

Kreps, David M. and Scheinkman, Jose A., "Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes," *Bell Journal of Economics*, Autumn 1983, *14*, 326–37.

Mankiw, N. Gregory and Whinston, Michael D., "Free Entry and Social Inefficiency," *Rand Journal of Economics*, Spring 1986, *17*, 48–58.

Matutes, Carmen and Regibeau, Pierre, "Mix and Match: Product Compatibility Without Network Externalities," mimeo., 1986.

McAfee, R. Preston, McMillan, John and Whinston, Michael D., "Multiproduct Monopoly, Commodity Bundling, and Correlation of Values," *Quarterly Journal of Economics*, May 1989, *104*, 371–84.

Ordover, Janusz A. and Saloner, Garth, "Predation, Monopolization, and Antitrust," in R. Schmalensee and R. D. Willig, eds., *The Handbook of Industrial Organization*, Amsterdam: North-Holland, forthcoming.

______, Sykes, A. O. and Willig, R. D., "Nonprice Anticompetitive Behavior by Dominant Firms Toward the Producers of Complementary Products," in F. M. Fisher, ed., *Antitrust and Regulation: Essays in Memory of John J. McGowan*, Cambridge, MA: MIT Press, 1985.

Posner, Richard A., *Antitrust Law: An Economic Perspective*, Chicago: University of Chicago Press, 1976.

Salop, S., "Monopolistic Competition with Outside Goods," *Bell Journal of Economics*, Spring 1979, *10*, 141–56.

Schmalensee, Richard, "Commodity Bundling by Single-Product Monopolies," *Journal of Law and Economics*, April 1982, *25*, 67–71.

______, "Antitrust and the New Industrial Economics," *American Economic Review*, May 1982, *72*, 24–28.

Selten, R., "A Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, *4*, 25–55.

Shaked, Avner and Sutton, John, "Relaxing Price Competition Through Product Differentiation," *Review of Economic Studies*, January 1982, *49*, 3–13.

Spence, A. Michael, "Product Selection, Fixed Costs, and Monopolistic Competition," *Review of Economic Studies*, 1976, *43*, 217–35.

Tirole, Jean, *The Theory of Industrial Organization*, Cambridge, MA: MIT Press, 1988.

Warren-Boulton, F. R., *Vertical Control of Markets: Business and Labor Practices*, Cambridge, MA: Ballinger, 1978.

Whinston, Michael D., "Tying, Foreclosure, and Exclusion," Harvard Institute of Economic Research Discussion Paper No. 1343, 1987.

U.S. Department of Justice, "Vertical Restraints Guidelines," *Antitrust and Trade Regulation Report*, Washington: Bureau of National Affairs, 1985.

The Economics of Product Patents

By MICHAEL WATERSON*

This paper develops a model of product patents in which the patent changes the nature of market entry behavior rather than preventing entry entirely. The model incorporates three levels of action, namely, the innovator's patenting decision, the potential entrant's location decision, and possible court action. The analysis demonstrates how the characteristics of the patents system and the enforcement framework can influence rivals' variety choices and thus the market equilibrium. It also considers how the system can in principle be adjusted to improve social welfare. (JEL *621*)

The purpose of this paper is to reexamine the question—what does a patent do? Certainly it acts as a deterrent to entry, as the established literature suggests,[1] but more precisely a patent on a product makes entry into the market by "close" substitutes costly. It prevents (or discourages) anything that is sufficiently close from being produced but does not foreclose a whole area in most cases. One might think of the case of a new child's toy, a new car safety restraint, or a new type of bottle opener. Here there is potential competition between the product that is patented (that suits some people very well, others not so well) and other products in this area. Thus we argue that the main impact of a product patent is not to create a monopoly but rather to affect the variety choices that rivals make. Moreover, the particular impact on variety choices is heavily influenced by the particular legal mechanisms that are used to enforce patent rights. The paper provides a simple framework for analyzing these variety issues, specifically addressing how patents affect the equilibrium number and nature of products produced, and whether it is socially desirable that they have these effects.

One major feature of the present paper is that the patent issued provides incomplete coverage for the patentee's idea in a specific sense—coverage is limited but inexact, and it can in principle be tested in the courts. In an important complementary theoretical analysis, Ignatius Horstmann, Glenn MacDonald, and Alan Slivinski (1985) assume limited but *exact* coverage.[2] In the jargon of patent law, they assume a very strict "fencepost" rather than a "signpost" system of patents. With a true fencepost system, there would be no need to refer to the courts over questions of interpretation, but in practice, court cases occur. And, as

*Department of Economics, University of Reading, Reading, RG6 2AA, England. I should like to thank Yannis Katsoulakos, Katrien Kesteloot, Sumner La Croix, Ian Molho, Norman Strong, participants at the 13th annual meeting of the European Association for Research in Industrial Economics in Berlin, at an ESRC Industrial Economics Study Group at the University of Sussex, at staff seminars at the Universities of Manchester, Melbourne, and York, Murdoch University Western Australia, the AGSM at the University of New South Wales, and the RSSS at the Australian National University and, last but by no means least, the *Review's* anonymous referees, for helpful comments and suggestions at various stages of preparation of this paper.

[1] The classic welfare analysis of patents appears in W. D. Nordhaus (1969). Some more recent studies have adopted a deterrence framework, in which patents provide a specific perpetual monopoly (see, for example, Richard Gilbert and David Newbery, 1982; Jennifer Reinganum, 1983; John Vickers, 1985), but these papers, and also those considering "patent race" issues (for example, Gene Grossman and Carl Shapiro, 1987), are largely concerned with the evolution of concentration over time and the question of whether monopolists persist.

[2] There are also some links between this paper and a recent welfare analysis of patent width by Paul Klemperer (1990).

Richard Levin (1986, p. 199) says, "There is no theoretical presumption that improving appropriability is desirable." This is one of the issues we examine below within our more fully structured version of the legal process. Our modeling approach enables predictions regarding changes in the legal system to be generated.

We choose to analyze the case of product patents because the idea of imperfect appropriability arises naturally in a product setting and because it enables us to raise interesting and largely unexplored issues regarding the effects of patents upon product locations. Admittedly, not all product patents fit within the precise framework adopted here, but this point is explored in an appendix (available from the author on request). Also, despite the focus on product patents, at least some of the issues raised here carry over to the case of patents on processes.

Section I sets out the model and develops its positive predictions. Section II develops some social welfare points through the medium of a specific example, and Section III contains a few concluding remarks.

I. The Model

Our model of patenting, like those cited in fn. 1, is rather stylized in its modeling of the inventive process. Firms are aware of what they will end up with; there is little uncertainty about the technology or demand.[3] All agents are assumed to maximize expected utility/profit arising from their actions. There is no dynamic learning process involved in producing new products. And the model largely abstracts from any strategic *pricing* behavior by the firms. We make this simplification in order to focus on strategic considerations of greater immediate interest. Finally, for simplicity, we have only two firms in the game, each of which produces only one product. Thus, our game is perhaps best thought of as a microcosm of a broader contest in which several firms compete, each with an array of products.

We make the following specific assumptions.

(i) The products in question are at particular points on a one-dimensional line of *finite* length $2x$, with customers evenly spaced along the line. The farther a customer is from a product, the higher is the "delivered price." This is most directly thought of as users obtaining lower consumer surplus at any price (F. M. Scherer, 1979) but might also fit with the firm in question having to incur additional promotional costs in order to make more sales, or more extensive product development (different types or styles), for instance.

(ii) The gross profits obtained by the two firms are concave functions of the extent of the market that is served to each side, given price, and for simplicity, the prices set are such that the entire market will be served.

(iii) The form of the patent naturally bounds the total market area in question (previously latent demand) by the broadest claim of the potential patentee.[4] Thus the first firm, I, irreversibly selects a market area and a product location, y. Absent entry, it earns profits (gross of any patent fee) that are written $\Pi(y,2x-y)$, as a function of the market areas, y and $2x-y$ to each side, over which its product is sold. (Notice price is implicit in the profit function.)

(iv) Entry is assumed to involve sinking substantial costs,[5] so it is irreversible and therefore not lightly undertaken. Assuming the potential entrant E locates 2θ units to one side of I $(0 \leq \theta \leq y/2)$, by convention the left, and that both firms set the same price, then I and E sell to distances to left

[3] In relation to uncertainty, the Horstmann et al. paper is clearly an advance on most previous work.

[4] The structure of UK patent applications, for example, is that they commence with a broad claim, then successively narrow this down in subsequent claims, as they proceed to a finer definition of the precise article. The courts are more likely to hold narrow claims as valid, which is what prompts our modeling framework for the legal process developed below.

[5] In order to economize on notation, in this section we assume that entry involves sinking costs equal to those the incumbent has incurred. This is not a material assumption, and it is dropped in Section II.

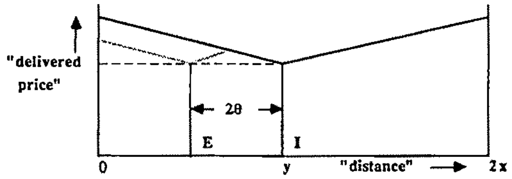


FIGURE 1. THE BASIC FRAMEWORK: I IS THE INNOVATOR AND E IS THE ENTRANT

and right of $(\theta, 2x - y)$ and $(y - 2\theta, \theta)$, respectively. Thus the scene in product space is as pictured in Figure 1.

(v) For much of the paper, we shall assume that, should firm I decide not to patent, the entrant will locate so that $\theta = 0$, as a result of the centripetal tendencies created by the finite market length. (This assumption is reexamined later.)

(vi) If the first firm patents, it incurs a nontrivial cost p (essentially, the legal fees plus a patent agent's charges), but benefits from the fact that $\theta > 0$. Since no further entry is envisaged, the potential entrant need not patent.

(vii) If I takes E to court over violation of patent rights, there are probabilities that I will win and E will lose with given expenditures on representation in court, and associated with the outcomes are payoffs. Following the action, both may remain in the market, or alternatively only I may actually be allowed to continue producing. However, it is still sensible to think of the payoffs in expected value terms, given risk neutrality on the part of the players. To put some structure onto this, we assume the incumbent receives expected net benefits of $C_i(\theta) = A(\theta) - K_i(\theta)$ from court action, while the entrant expects $C_e(\theta) = -A(\theta) - K_e(\theta)(< 0)$, A being a nonnegative transfer of damages and K_i and K_e being deadweight legal costs to the parties. Then the respective payoffs are

$$[\Pi(\theta, x) - p + C_i(\theta),$$
$$\Pi(x - 2\theta, \theta) + C_e(\theta)].$$

It is natural to assume that $C_i' < 0$ with $C_i(\theta) > 0$ for θ small, but $C_i(\theta) < 0$ for large θ due to negligible damages in this case. It also seems reasonable to assert that $C_e' > 0, C_e'' < 0, C_e(0) < -\Pi(x, 0)$; that is, court costs increase at an increasing rate as E moves toward $I(\theta \to 0)$, and direct copying is unprofitable.

The sequence of events is as follows. Firm I selects its location and decides whether to patent. Firm E then decides whether, and if so where, to enter. If I did not patent and E wishes to enter, then I cannot prevent entry. If I has patented, then after E has entered, I has the option of taking E to court over patent violation. It will save time if we note at this stage that firm I's equilibrium location is at x, the center of the market,[6] and analyze subsequent decisions with I's location given.

In analyzing I's and E's decisions based upon the model outlined above, we utilize the now-standard subgame perfect equilibrium concept. The game tree corresponding to this model is represented in extensive form in Figure 2. It is clear that there is more than one potential equilibrium outcome. In fact, the only possibility that can be ruled out without further consideration is that where I does not patent but E stays out, since E does better by entering knowing that I will accommodate it, and any threat by I will not be credible in this case.[7]

However, we can proceed further by making use of the assumptions, in particular (vii). Simple functions consistent with these properties are illustrated in Figure 3. These enable us to determine both θ and the range of possible equilibria. A maximum on $[\Pi(x - 2\theta, \theta) + C_e(\theta)]$ occurs at a value of θ^*, point C in the figure, representing E's best position if it expects a court battle. E will only expect this if $C_i(\theta) > 0$; the threat to fight must be believable (Dixit, 1982; Steven Salop, 1979). E's alternative is to

[6] Observe that if I were to locate at $x + \Delta$, rather than x, then without entry, its profits would be $\Pi(x + \Delta, x - \Delta) < \Pi(x, x)$, and with entry, $\Pi(\theta, x - \Delta) < \Pi(\theta, x)$.

[7] Thus for example we might allow firm I to threaten to price down to marginal cost, but since this involves both I and E in losses, it is not credible (see, for example, Avinash Dixit, 1982).

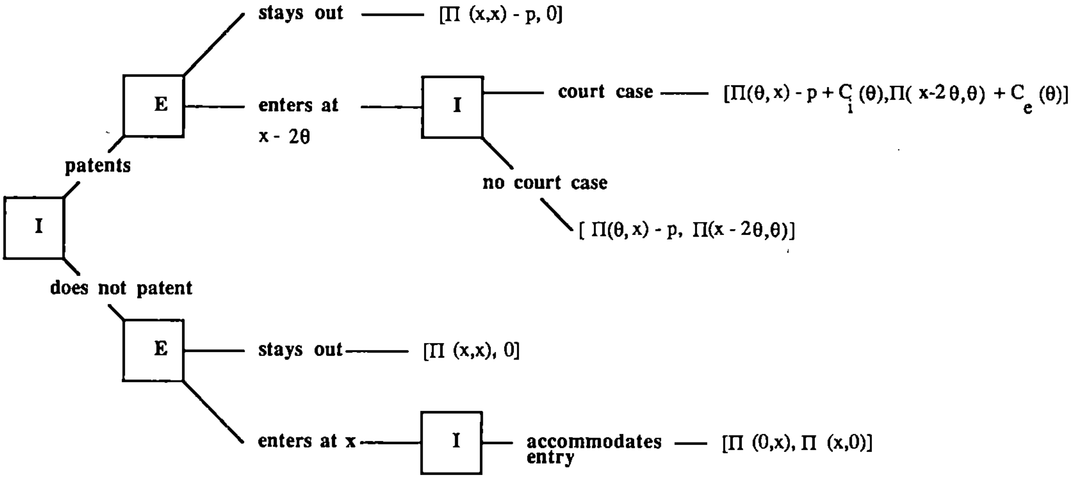


Figure 2. The Game In Extensive Form

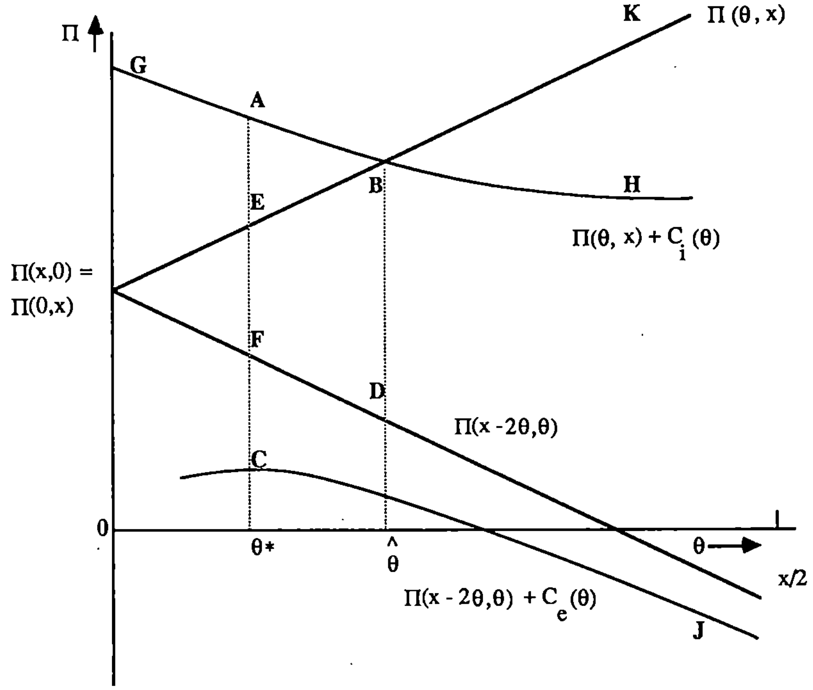


Figure 3. The Profit Functions

choose a θ value, here called $\hat{\theta}$, which gives the highest profits E can expect conditional on not triggering a court action, that is, where $C_i(\hat{\theta}) = 0$. This yields E profits of $\Pi(x - 2\hat{\theta}, \hat{\theta})$. As the figure is drawn, $\Pi(x - 2\hat{\theta}, \hat{\theta}) > \Pi(x - 2\theta^*, \theta^*) + C_e(\theta^*)$ (point D is higher than C), meaning that should I patent, the best E can do in the subgame in which it decides to enter is to choose the variety $\hat{\theta}$. But it is clear that the opposite could be true. Notice also that $AE < CF$; in the event of a court action, there is a net loss to the pair of firms, representing their legal costs.

TABLE 1—EQUILIBRIUM OUTCOMES OF THE GAME

Relationship Between Profit Values	Outcome	Text Reference
$\Pi(x-2\hat{\theta},\hat{\theta}) > \Pi(x-2\theta^*,\theta^*)+C_e(\theta^*)$, that is, $\hat{\theta}$ Relevant		
$\Pi(x-2\hat{\theta},\hat{\theta}) > 0;\ \Pi(\hat{\theta},x)-p > \Pi(0,x)$	Patent; No Court Case	(1)
$\Pi(\hat{\theta},x)-p < \Pi(0,x)$	No Patent; No Court Case But Entry	(2)
$\Pi(x-2\hat{\theta},\hat{\theta}) < 0;\ \Pi(\hat{\theta},x)-p \gtrless \Pi(0,x)$	Patent; E Stays Out	(3)
$\Pi(x-2\hat{\theta},\hat{\theta}) < \Pi(x-2\theta^*,\theta^*)+C_e(\theta^*)$, that is, θ^* Relevant		
$\Pi(x-2\theta^*,\theta^*)+C_e(\theta^*) > 0, \Pi(\theta^*,x)-p+C_i(\theta^*) > \Pi(0,x)$	Patent, But Court Case	(4)
$\Pi(\theta^*,x)-p+C_i(\theta^*) < \Pi(0,x)$	No Patent; No Court Case But Entry	(2)
$\Pi(x-2\theta^*,\theta^*)+C_e(\theta^*) < 0, \Pi(\theta^*,x)-p+C_i(\theta^*) \gtrless \Pi(0,x)$	Patent; E Stays Out	(3)

The various possible situations arising from Figure 2 are written down in Table 1. Several features of interest emerge from these calculations. Most obviously, if the best E can do given the patent is to make a loss, E cannot credibly threaten to enter. Nevertheless I will patent, as long as we assume that $\Pi(x,x)-p > \Pi(0,x)$, a most likely occurrence at current fee levels. Therefore equilibrium 3 results. However, if E can make a profit, any one of the other three equilibria can arise. Obviously, if I is better off in terms of final profits not patenting, then equilibrium 2 results. Given the current sizes of patent fees though, it is more likely that I is better off patenting, in which case the two firms may implicitly agree not to fight (equilibrium 1)[8] or, interestingly, they both may actually be better off going to court (equilibrium 4). This last outcome may be viewed as an example of a curious type of prisoners' dilemma. The parties jointly would be better off at $\hat{\theta}$ with profit-sharing but E has an individual interest in being at θ^* even in the knowledge of the court case, while location at θ^* means I has an incentive to bring the case.

Thus a variety of equilibria are possible, with the specific outcome dependent upon characteristics of the patent system itself.

Let us now consider some comparative static properties of the patents system and its associated legal apparatus. Note first that a reduction in the deadweight legal costs involved in court action will raise curve CJ in Figure 3 as well as raising $GABH$. These movements lead to some straightforward effects but also to others less obvious. One effect is to shift $\hat{\theta}$ to the right. However, it also leads to an increased likelihood of the θ^* equilibrium.[9] Point C in Figure 3 is raised, at the same time as D shifts down line FD. Thus streamlining court procedure makes the potential entrant more wary of introducing a close substitute for I's product, but also releases some of the latent demand for court action.

So far, it has been assumed that, even if the legal system may award costs (as it can in the UK), the costs of engaging in the action far exceed those incurred in, and therefore in principle recoverable through, the courts. But consider the effect of introducing a purely contingent-fee based legal cost framework (rather unrealistically, ignoring unrecoverable costs). In that case, $C_i(\theta) = A(\theta) - K_i(\theta) > 0$. Therefore curve

[8] Notice that between $\theta = 0$ and $\hat{\theta}, \Pi(\theta,x)+C_i(\theta) > \Pi(\theta,x)$, which is greater than the profits earned over the same distance with price competition, so that price competition by I as a response to entry is never credible in a single-shot game. Moreover, price competition would make entry, at $\theta = 0$ at least, unprofitable for E. Hence, we ignore this possibility.

[9] It is fairly clear that, under a wide range of circumstances, the $\hat{\theta}$ set of equilibria is socially to be preferred to the θ^* set.

GABH lies entirely above line *EBK*. In consequence, the $\hat{\theta}$ equilibrium disappears, so that entry, if it occurs, must be at the edge of the market, or a court case will necessarily ensue.

Notice that changes in the magnitude of p leave both $\hat{\theta}$ (and θ^*) equilibria unchanged, since lines *GABH* and *EBK* are net of patenting costs. Thus the only direct effect of a reduction in patent fees is to increase the likelihood of patenting. However, if we make a mild additional assumption, we may say something rather more interesting regarding the propensity to patent.

To do this, consider the hitherto neglected possibility that the desire for variety is sufficiently strong for *E*'s most profitable position to be some way from *I*. Return to the "normal" penalty structure situation, where $C_i(x/2) < 0 < C_i(0)$ and $A'(\theta) < 0$, and assume $\hat{\theta}$ and not θ^* is relevant. In that case, *I*'s propensity to patent will be lower when customers' desire for variety is strong. This transpires because, under the stated assumptions, *E*'s optimal value of θ in the absence of a patent, θ^n, say, is positive, which implies *I* would find itself some way up the line $\Pi(\theta, x)$. However, there would be no reason for $\hat{\theta}$ to be different. *I*'s comparison in deciding whether or not to patent is between $\Pi(\theta^n, x)$, which increases with θ^n, and $\Pi(\hat{\theta}, x) - p$. Therefore as the desire for variety, and so θ^n, increases, the former value is increasingly likely to be the larger, so leading *I* to choose not to patent.

Casual empiricism is perhaps dangerous, but interindustry comparisons of the ratio between the proportion of firms with patents and the proportion with R&D programs, calculated from data reported in Zvi Griliches, Ariel Pakes, and Bronwyn Hall (1987), are suggestive. The lowest propensity to patent, defined in this sense, comes from the industry group "office, computing and accounting equipment," and low values are also featured in "communication equipment" and "professional and scientific equipment." Goods like "primary metals" and "petroleum refining" have very high values. Each of the low propensity groups could be considered as cases where customers' desire for variety is very strong, and vice versa.

The outcomes in our model may also be compared with those arising in Horstmann et al.'s rather different model. Like us, they predict that not all innovations will be patented, but they find those that are not patented will be imitated, whereas our model suggests they will often be copied. We would argue that at least some observations support our version (for example, drugs in Italy, pirating of computer software). Their third major prediction is that patented goods or processes will be neither imitated nor duplicated. This rather counterintuitive outcome arises because in their model there is some positive probability that imitation is unprofitable, and the occurrence of a patent signals a case where the entrant's expected profit is negative; they make no claims for their prediction in those cases where imitation or duplication would always be profitable in the absence of a patent but, given the patent, imitation need not be. In summary, the deterrence effect of patents arises in very different ways in their framework and ours.

II. A Specific Illustration

This section describes a simple example that illustrates a number of results in the preceding section but that focuses upon the welfare effects of location and firm numbers raised by the model. Since duopoly frameworks commonly will involve prices lower than under monopoly, whereas here by nature of the example prices are parametric, this is not an innocuous simplification. Our justification is that it enables us to concentrate upon issues of current interest regarding numbers and locations of products and moreover that, by contrast with effects arising from proportionately large changes in product numbers, deadweight welfare loss is a second-order effect.

The example uses a formulation popularized by Curtis Eaton and Richard Lipsey (1978) among others—though we differ in employing the assumption of a bounded market, given the nature of the patent application. Consumers at some point in the product space represented in Figure 1 have an (indirect) utility function:

$$(1)\qquad V = e^{-P_c/P_0} + I/P_0,$$

where

$$(2)\qquad P_c = P_f + zt,$$

P_f being the price set by the firm selling the nearest product, t the unit cost associated with that product being a distance z from the consumer's ideal, I income and P_0 the price of the numeraire good, set equal to unity. All consumers are alike save in their views regarding the best product, and they are uniformly distributed across the market. Each firm is assumed to face costs

$$(3)\qquad C_j = cQ + F_j;\ j = i, e$$

linear in total output Q, with a fixed element F_j that may be smaller for firm E than for firm I. Deriving demand from (1) and integrating across area served yields profits, before deduction of any patenting or legal costs, as

$$(4)\quad \Pi_j = (P_f - c)\frac{e^{-P_f}}{t} \times [2 - e^{-D_L t} - e^{-D_R t}] - F_j,$$

D_L and D_R being the left- and right-hand limits to sales. By differentiation, optimal price assuming price matching is

$$(5)\qquad P_f^* = 1 + c.$$

Using (1) and (5), consumer surplus generated over the area covered by a specific firm operating D_L units to its left and D_R units to its right is

$$(6)\quad CS = \frac{e^{-(1+c)}}{t}[2 - e^{-D_L t} - e^{-D_R t}].$$

Social surplus arising from the production of a particular product is $S = CS + \Pi$, obtainable from (4) and (6). Thus we write down the surplus created when one centrally placed firm I, begins operations as

$$(7)\quad S_i(x,x) = \frac{2e^{-(1+c)}}{t}[2 - 2e^{-xt}] - F_i.$$

Suppose now that there are two firms, incumbent and entrant. The entrant has to decide where it would be best to come in. Using (4) its profit is

$$(8)\quad \Pi_e(x - 2\theta, \theta) = \frac{1}{t}e^{-(1+c)} \times [2 - e^{-(x-2\theta)t} - e^{-\theta t}] - F_e.$$

Assuming there is no patent constraining location, this is optimized for a value of θ given by

$$(9)\qquad \theta_o = \mathrm{Max}\left\{\frac{x}{3} - \frac{1}{3t}\ln 2, 0\right\}.$$

This implies locating right next to the incumbent unless the "transport cost" exceeds approximately 0.7 per unit distance x.[10]

[10] It may be useful to illustrate the possibility that a θ^* equilibrium may arise, within the context of this specific example. Assume

$$\begin{aligned} C_i(\theta) &= \Pi_e(x-2\theta,\theta) - k & \theta &< 0.1\\ &= 0.1\Pi_e(x-2\theta,\theta) - k & \theta &\geq 0.1\\ C_e(\theta) &= -\Pi_e(x-2\theta,\theta) - k & \theta &< 0.1\\ &= -0.1\Pi_e(x-2\theta,\theta) - k & \theta &\geq 0.1 \end{aligned}$$

and consider the following parameter values: $x = c = 1, F_e = 0.01, t = 0.1, k = 0.0083$. Then $\hat{\theta} = 0.3$, at which point E's net profits are 0.083. θ^* is 0.1, clearly, at which point E's net profits are greater, at 0.0885.

With both firms in the market, the general expression for social surplus is

$$(10)\quad S(\theta,x;x-2\theta,\theta) = \frac{2e^{-(1+c)}}{t} \times [4-2e^{-\theta t}-e^{-xt}-e^{-(x-2\theta)t}] - F_i - F_e.$$

Here it is easily confirmed that the socially optimal value for θ is $x/3$. Hence by comparison with (9), there is always a centripetal tendency in this model. Thus there are two potential sources of static social welfare loss: having the wrong number of incumbents and having the wrong array of products.

Assume for the time being that t is sufficiently low for E, left to itself, to desire a location next to I. Then it is immediately seen, by comparing (10) with $\theta = 0$, with (7), that society would be better off in a static sense with a monopolist, for any $F_e > 0$. This also provides I with more profit, and hence more incentive, than does freer entry, since $\Pi_i(x,x) > \Pi_i(\theta,x)$. We have:

PROPOSITION 1: *There are circumstances where if entry can be prevented by use of a patent system, the innovator will receive more (dynamic) encouragement in the form of expected profits as well as society being better off in the sense of having larger social surplus.*

When we introduce the possibility of the incumbent patenting, then the entrant's location will never optimally be at $\theta = 0$, assuming there are penalties for transgressing the law. Where $\theta > 0$, it is possible that society benefits from having two suppliers. The patent, as we shall see, increases this tendency.

In society's eyes, it is worthwhile having two firms rather than one if $S(\theta,x;x-2\theta,\theta) > S_i(x,x)$ (i.e., the value of (10) is greater than the value of (7)). From the viewpoint of the potential entrant, it is worthwhile entering if the value of $\Pi_e(x-2\theta,\theta)$, that is, (8), is positive. It is clear that E may enter when entry is socially undesirable since there will be a "business stealing" effect (Gregory Mankiw and Michael Whinston, 1986), which may or may not be outweighted by the gain in variety. The key factors here are the size of the entrant's fixed costs, consumers' desire for variety, and the form of the penalty scheme.

Consider first a stylized "signpost" system of patents that does not designate any particular product, save an exact copy, as outlawed, but that involves an increasing expected value of penalty as the entrant product impinges more strongly upon the patented one.[11] Specifically, take two versions:

A. A penalty on E's profits that is proportional to closeness, varying between no penalty at $\theta = x/2$ and 100 percent penalty at $\theta = 0$, namely, $(1-2\theta/x)\times \Pi_e(x-2\theta,\theta)$. Alternatively, this may be thought of as a penalty of $\Pi_e(x-2\theta,\theta)$ with probability $(1-2\theta/x)$

B. A penalty equal to I's loss in profit $\Pi_i(x,x)-\Pi_i(\theta,x)$, imposed with probability $(1-2\theta/x)$

These are stylized versions of the two common forms of damages in patent claims, namely, breach of contract (scheme A, second interpretation) and tort (scheme B). They may be contrasted with a perfect "fencepost" system that bars entry within a complete specified area, say, up to the edge of the market (with the penalty being a fine equal to the profits made by the entrants), and with blockaded entry.

In order to analyze the relative welfare effects of these schemes, we adopt a numerical approach.[12] Some examples are given in Table 2.

[11] As W. R. Cornish (1981, p. 128) puts it, "to some extent it is left to the court to work out the proper scope of the monopoly from the description of the invention."

[12] Details of the calculations involved are omitted, as they are straightforward but tedious. For simplicity in calculating social welfare, legal costs are assumed to be lump-sum transfers to lawyers, and patenting costs are ignored.

TABLE 2—SOME SIMULATION RESULTS

				Blockaded	No Penalty		Scheme A		Scheme B		Fencepost ($\theta = 0.5$)
Case	t	F_i	F_e	S	θ_0	S	θ	S	θ	S	S
1	0.05	0.05	0.05	0.4780	0	0.4280	0.32	0.4324	0.48	0.4316	0.4313
2	0.05	0.1	0.05	0.4280	0	0.3780	0.32	0.3824	0.48	0.3816	0.3813
3	0.05	0.1	0.1	0.4280	0	0.3280	0.13	0.3308	N	0.4280	N 0.4280
4	0.5	0.03	0	0.3960	0	0.3960	0.45	0.4276	0.40	0.4308	0.4225
5	0.5	0.03	0.03	0.3960	0	0.3660	0.38	0.4016	0.40	0.4008	0.3925
6	0.5	0.05	0.03	0.3760	0	0.3460	0.38	0.3816	0.40	0.3808	0.3725
7	0.5	0.05	0.05	0.3760	0	0.3260	0.32	0.3623	0.40	0.3608	0.3525
8	0.5	0.1	0.05	0.3260	0	0.2760	0.32	0.3123	0.40	0.3108	0.3025
9	0.5	0.1	0.1	0.3260	0	0.2260	0.08	0.2411	N	0.3260	N 0.3260
10	1	0.03	0.03	0.3122	0.10	0.3117	0.37	0.3405	0.37	0.3405	0.3241
11	1	0.05	0.05	0.2922	0.10	0.2717	0.32	0.3012	0.40	0.2986	0.2841
12	1	0.1	0.05	0.2422	0.10	0.2217	0.32	0.2518	0.37	0.2505	0.2341
13	1	0.1	0.1	0.2422	N	0.2422	N	0.2422	N	0.2422	N 0.2422
14	2	0.05	0.05	0.1840	0.22	0.2046	0.31	0.2141	0.34	0.2145	0.1881
15	4	0.05	0.05	0.0829	0.28	0.1136	0.28	0.1136	N	0.0829	N 0.0829
16	4	0.1	0.05	0.0329	0.28	0.0636	0.28	0.0636	N	0.0329	N 0.0329
17	4	0.1	0.1	0.0329	N	0.0329	N	0.0329	N	0.0329	N 0.0329
18	5	0.05	0.05	0.0575	0.29	0.0841	N	0.0575	N	0.0575	N 0.0575

Notes: $c = x = 1$ throughout.
S = social welfare, calculated from equation (10).
N = entry does not occur; monopolist remains.

Various points of note emerge from the table. Most obviously, the patent system, like free entry, commonly allows the second product to achieve some position in the market rather than barring it completely. Particularly for the lower t values, where consumers value variety only moderately, social welfare is markedly improved thereby over the free entry (and, sometimes, the blockaded entry) position. We have:

PROPOSITION 2: *The presence of a patents system can lead to products being better located in economic space than in its absence.*

For very low t values, where consumers care little about variety, Proposition 1 dominates and blockaded entry is usually better, of course.

It is also noticeable that, once we move toward rather high values for t (for example, $t = 4$ per unit distance), this effect is reduced since a substantial separation between products would exist independent of the patent system, and any penalty imposed is as likely to discourage entry as it is to improve separation. In particular, a rather limited degree of protection, if any, seems desirable once we move to high t values, as cases 10–18 show.

Generally, the fencepost system produces over-much separation of the products while permitting entry and so performs poorly over the range of values listed in the table. Scheme B appears the most resilient to parameter changes in the sense that it always does better than either scheme A or the fencepost system (or both) for the range of values given. This is interesting since scheme B can be thought of as a proportional tax on (privately but not socially desirable) business stealing. Scheme A performs well except when F_e is large, in which circumstance the penalty will be small since E's profits are low.

III. Concluding Remarks

This paper has pointed out the very complex effects of a patent system. Giving a monopoly right can often prove socially very worthwhile both in terms of encouraging firms to design innovative products and in terms of providing society with the right number of products. However, there remain the very important questions of deciding how extensive the right should be, and designing the system by which those rights are exercised. The positive desire for variety in some industries, coupled with the influence of the patent on the extent of variety that results, seems a noteworthy addition to the range of relevant considerations in a patent system. So also does the structural point that a contingent fee system for patent court cases might well largely destroy the potential gains from increased variety.

This paper, like many others concerned with welfare aspects of patents, faces one major problem in drawing conclusions: it is hardly likely that patent law can be made industry specific. Yet one implication may still be drawn. Broadly, patents provide wide-band protection as a result of considerable effort in formalization and, if necessary, defense on the patentee's part, whereas copyright provides easier but more narrowly defined protection. In some product areas, (for example, computer software), there is debate as to which is the more appropriate system since it is not clear whether one or the other should apply. Here, that feature of the model, which tentatively suggests that where variety is valued very highly, broad band protection is likely to be socially inferior, would imply that copyright is the more appropriate form of protection for those cases where consumers desire a plethora of specific applications for specific situations.

REFERENCES

Cornish, W. R., *Intellectual Property: Patents, Copyright, Trademarks and Allied Rights*, London: Sweet and Maxwell, 1981.

Dixit, Avinash K., "Recent Developments in Oligopoly Theory," *American Economic Review*, May 1982, *72*, 12–17.

Eaton, B. Curtis and Lipsey, Richard G., "Freedom of Entry and the Existence of Pure Profit," *Economic Journal*, September 1978, *88*, 455–69.

Gilbert, Richard J. and Newbery, David M. G., "Pre-emptive Patenting and the Persistence of Monopoly," *American Economic Review*, September 1982, *72*, 514–26.

Griliches, Zvi, Pakes, Ariel and Hall, Bronwyn, "The Value of Patents as Indicators of Inventive Activity," in Partha Dasgupta and Paul Stoneman, eds., *Economic Policy and Technological Performance*, Cambridge: Cambridge University Press, 1987, ch. 4.

Grossman, Gene M. and Shapiro, Carl, "Dynamic R&D Competition," *Economic Journal*, June 1987, *97*, 372–87.

Horstmann, Ignatius, MacDonald, Glenn M. and Slivinski, Alan, "Patents as Information Transfer Mechanisms: To Patent or (Maybe) Not to Patent," *Journal of Political Economy*, October 1985, *93*, 837–58.

Klemperer, Paul, "How Broad Should the Scope of Patent Protection Be?"*Rand Journal of Economics*, forthcoming, 1990.

Levin, Richard C., "A New Look at the Patent System," *American Economic Review*, May 1986, *76*, 199–202.

Mankiw, N. Gregory and Whinston, Michael D., "Free Entry and Social Inefficiency," *Rand Journal of Economics*, Spring 1986, *17*, 48–58.

Nordhaus, W. D., *Invention, Growth and Welfare*, Cambridge, MA: MIT Press, 1969.

Reinganum, Jennifer F., "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, September 1983, *73*, 741–78.

Salop, Steven C., "Strategic Entry Deterrence," *American Economic Review*, May 1979, *69*, 335–38.

Scherer, F. M., "The Welfare Economics of Product Variety: An Application to the Ready-to-Eat Cereals Industry," *Journal of Industrial Economics*, December 1979, *28*, 113–34.

Vickers, John, "Pre-emptive Patenting, Joint Ventures, and the Persistence of Oligopoly," *International Journal of Industrial Organization*, September 1985, *3*, 261–73.

Are Treble Damages Neutral? Sequential Equilibrium and Private Antitrust Enforcement

By DAVID BESANKO AND DANIEL F. SPULBER*

A sequential equilibrium model of private antitrust enforcement is presented. Consumers have incomplete information about cartel costs and cannot accurately estimate a priori the damage recovery from an antitrust action. Consumers are able to infer cartel costs from the equilibrium pricing strategy of firms. The universal divinity criterion is used to characterize the sequential equilibrium. It is shown that for a sufficiently large damage multiple, antitrust enforcement effectively increases social welfare. (JEL *026, 612)*

The dominant form of antitrust enforcement in the United States is the private party lawsuit. In the 1980s private lawsuits outnumbered public lawsuits by over 10 to 1 (Steven Salop and Lawrence White, 1988). The large volume of private antitrust activity has engendered heated debates in recent years regarding the welfare and efficiency effects of antitrust law.[1] Richard Posner (1976, p. 35) observes that the number of private actions has "grown explosively," and he expresses "concern about the overexpansion of the antitrust laws." Some research has raised the possibility that private antitrust enforcement is neutral, with no deterrent effects on competitive behavior. This would suggest that the antitrust laws are ineffective, and the large legal costs of private enforcement constitute a significant deadweight welfare loss. While not attempting to resolve the ongoing antitrust policy debate, the present paper provides a game-theoretic model of private antitrust that addresses the issue of neutrality. It is shown that in a sequential equilibrium with asymmetric information, private antitrust enforcement can increase social welfare if the probability of conviction is sufficiently large.

It is widely recognized that the treble damages remedy can create "perverse incentives" for private enforcers. William Breit and Kenneth Elzinga (1974, 1985) have argued that under a multiple damages remedy, buyers may have no incentive to avoid the damage of monopolistic overcharges by searching for competitive substitutes. Indeed, buyers may even seek to increase their damages by increasing their purchases from a price-fixing cartel.[2] Frank Easterbrook (1985, p. 451) notes that to the extent that buyers "have perfect information and enforcement is costless, they view the future recovery as a cents-off coupon attached to each purchase." Recognizing these incentives, Jonathan Baker (1988) and Steven Salant (1987) analyze models of private antitrust enforcement and find that the multiple damages remedy has neutral welfare consequences. In their models, consumers increase demand in anticipation of collecting treble damages, and this offsets the increase in the cartel's price due to the expected penalty. As a result, the cartel's output exactly equals the unconstrained monopoly level. Consumers' and producers'

*Department of Business Economics and Public Policy, Indiana University; and Kellogg Graduate School of Management, Northwestern University. We would like to acknowledge the helpful comments and suggestions of Jeffrey Banks and Jonathan Baker. Responsibility for any remaining errors is ours.

[1] William Breit and Kenneth Elzinga (1985, p. 405) observe that the antitrust literature and congressional debates have been "almost uniformly critical" of private treble-damages actions.

[2] Elzinga and Breit (1976) suggest that these incentives may have been at work in the electrical equipment conspiracy of the 1950s.

surpluses are unchanged because the expected damage award is a pure transfer.[3]

The neutrality of various legal institutions has often been stressed in the law and economics literature. The classic analysis of Ronald Coase (1960) argued that the outcome of private bargaining over an externality was unaffected by whether property rights were assigned to the polluter or to the victim.[4] As Coase stressed, the neutrality of legal institutions depends upon the absence of transaction costs, including availability of full information. Private antitrust enforcement differs from enforcement of criminal law (see David Besanko and Daniel Spulber, 1989b) in that the two parties are in a contractual relationship. Private antitrust enforcement is similar in this sense to actions under contract or tort law that involve a buyer and a seller. Transaction costs can affect contractual terms between cartel members and their customers and thus determine the effectiveness of private antitrust. The purpose of the present analysis is to examine the effects of imperfect information on the neutrality of private antitrust enforcement.

This paper presents a model of private antitrust enforcement in which the cartel has better information about its production costs than prospective buyers. Consumers thus make purchase decisions with imperfect information about whether price-fixing has occurred and the extent of the antitrust violation (the markup of price over marginal cost). A sequential equilibrium is obtained in a game-theoretic model in which the cartel sets a price and consumers then decide how much to purchase and whether to take legal action against the cartel. The universal divinity criterion due to Jeffrey Banks and Joel Sobel (1987) is used to restrict consumers' "off the equilibrium path" beliefs. In equilibrium, there is a positive relationship between the cartel's price and consumers' expectations of the cartel's costs. In contrast to the situation in which consumers have full information about the cartel's cost, a higher price will not necessarily lead consumers to believe that the antitrust violation is more severe and thus to expect a larger damage award. Asymmetric information about the cartel's costs thus introduces a friction that offsets the neutrality due to the "perverse incentive." As a result, a multiple damages remedy can result in welfare improvements if the probability of conviction or damage multiple is sufficiently large.

The paper is organized as follows. The model of private antitrust enforcement is set out in Section I. Neutrality of private antitrust under full information is examined in Section II. The sequential equilibrium under asymmetric information is presented in Section III and further characterized in Section IV. Conclusions are in Section V.

I. A Model of Private Antitrust Enforcement

A group of identical firms in a given industry must decide whether or not to form a cartel. If a cartel is formed, the market price p is chosen cooperatively. Establishing the cartel is costless but carries the risk that the firms may be sued by consumers for price-fixing and the prospect of multiple damages. If the firms choose not to form a cartel, Bertrand competition ensues, and the market equilibrium price equals marginal cost. Besanko and Spulber (1989a) analyze a similar cartel model in a public enforcement setting.

The cartel's marginal cost can take values in the set $\{\theta_1,\dots,\theta_m\}$, where $\theta_i < \theta_{i+1}, i \in \{1,\dots,m-1\}$. Cartel costs are not known to consumers. Let γ_i be a consumer's prior probability that the cartel's cost is θ_i. The prior probability is the same for all consumers and is common knowledge.

If a lawsuit results in the cartel being found guilty of fixing price p above marginal cost, the courts are assumed to determine perfectly the overcharge $p-\theta$. The winning consumers collect damages $t(p-\theta)q$, where

[3] The "perverse incentive" also operates in Pablo Spiller's (1986) dynamic model in which buyers make repeated purchases from a cartel over time in anticipation of obtaining treble damages. Buyers stop making purchases and sue the cartel only when total damages reach the limits of the cartel's liability. Limited liability leads to an outcome that is not welfare neutral.

[4] See Spulber (1987) for a related analysis of damage remedies for breach of contract.

$t > 0$ is a damage multiple. Let $\beta \in (0,1)$, represent the probability that the cartel will be found guilty by the courts. This allows for legal rules that require evidence of collusion in addition to the existence of markups above marginal cost.[5] We assume that β is common knowledge to consumers and firms. If $\beta t \geq 1$, expected profits from cartelization would always be nonpositive. Thus, if the conviction probability or the damage multiple is sufficiently large, multiple damages result in total deterrence of price-fixing. Accordingly, it is assumed throughout that $\beta t < 1$. For example, with treble damages the probability of conviction of a guilty cartel must be less than one-third. The issue is then to determine whether private antitrust enforcement has marginal deterrence effects on industry behavior.

The cartel and consumers interact over two dates. At date zero, the firms decide whether to form a cartel. If they do, the market price p is chosen cooperatively and announced to consumers. Upon observing this price, consumers make a utility-maximizing purchase decision. Then, at date one, consumers can sue for price fixing.[6] A consumer compares the expected return from suit with the legal costs of suit. If legal costs are normalized to zero, then under a multiple damages remedy, the consumer will sue if and only if $p > \theta$. If the consumer knows θ, he can adjust his purchases to reflect the anticipated damage award. If cartel costs θ are unknown, the consumer's purchase level and decision to sue depend on his expectations about θ. The consumer's expectations will depend on the observed equilibrium cartel price.

The cartel is assumed to face a market of identical consumers. Industry demand at date zero thus can be derived from the maximization problem of a representative consumer. The representative consumer is assumed to have quasi-linear preferences $V(q)+x$ over the cartel's product q and a numeraire good x. The total willingness to pay $V(q)$ satisfies $V(0)=0$, $V'(q)>0$, $V''(q)<0$. Let $P(q) \equiv V'(q)$ and $Q(p) \equiv P^{-1}(p)$. In the absence of antitrust enforcement, $Q(p)$ would be the industry demand function, and $P(q)$ would be the industry inverse demand.

The setting we consider has two noteworthy features. First, notice that consumers cannot make demand decisions or choose possible legal actions until the market price has been observed. This implies that consumers cannot make commitments to a particular course of action. In this way the situation differs from principal agent–type models of regulation or antitrust (for example, Besanko and Spulber, 1989a) in which enforcement authorities can commit to a probability of suit so as to deter monopoly pricing. Public authorities can issue antitrust guidelines, invest agency funds, or target markets for investigation and prosecution, although public agencies can also encounter problems in making credible commitments; see Besanko and Spulber (1989b). Since private enforcement efforts are costly to implement and are often unobservable to violators, enforcement efforts are not credible threats. This suggests the present sequential equilibrium model in which consumers make purchase and enforcement decisions after firms have made price-fixing decisions.

Second, since industry costs are unknown to consumers, consumers may gain information about industry costs (and thus the extent of possible antitrust violations) from the market price. In equilibrium, a consumer's demand and decision to sue are based upon the information conveyed by prices. The information content of prices

[5] The probability of conviction is assumed to be independent of the seller's markup, and thus of the cost uncertainty. This is more restrictive than Baker (1988) and Salant (1987) who allow the probability of conviction to depend on cartel output. The joint probability of detection and conviction may increase with the markup. The present model makes detection exogenous. An endogenous conviction rate based on cartel behavior would be of interest for future work.

[6] It is assumed that consumers and the cartel have the same discount factor and that the discount factor is included in the damage multiple t. (That is, let $t = \delta t'$, where δ is the discount factor and t' is the nominal damage multiple.)

plays an important role in the cartel's pricing strategy.[7]

II. Neutrality Under Symmetric Information About Industry Costs

The neutrality of private antitrust enforcement in now examined. It is then shown that neutrality can be eliminated by various institutional constraints, even under full information. These constraints are relaxed in the next section to highlight the effects of asymmetric information.

Baker (1988) and Salant (1987) show that in the full-information setting with zero legal costs, private antitrust enforcement will have neutral welfare and distributional consequences. Their result can be derived as follows.

The consumer's utility maximization problem at date zero is

$$\max_{q \geq 0} \{V(q) - pq + \beta t(p - \theta) q I_{[\theta,\infty)}(p)\},$$

where $I_{[\theta,\infty)}(p)$ is an indicator function taking on value one for $p \geq \theta$ and a value of zero otherwise. If the expected damage multiple βt is less than one and price is at least as great as marginal cost, the consumer's demand $q^f(p,\theta)$ is given by

$$(1) \quad q^f(p,\theta) = Q((1-\beta t)p + \beta t\theta).$$

For future reference, note that $q^f(p,\theta)$ is decreasing in θ. That is, a consumer will purchase more from a cartel with lower production costs because at any given market price the consumer expects a greater damage award.

The consumer's *marginal willingness to pay* for the cartel's output under multiple damages is

$$p^f(q,\theta) \equiv [P(q) - \beta t\theta]/(1-\beta t).$$

For any quantity of output below the perfectly competitive level, a multiple damages award increases the consumer's marginal willingness to pay. This follows from the fact that with a multiple damages remedy, consumer marginal willingness to pay $p^f(q,\theta)$ exceeds the marginal benefit $P(q)$ from consumption.

If the cartel forms, the profit-maximizing price solves[8]

$$(2) \quad \max_{p \geq \theta_i} (p - \theta_i) q^f(p,\theta_i)(1-\beta t).$$

The solution is given by

$$p^f(\theta_i) = (p^M(\theta_i) - \beta t\theta)(1-\beta t)^{-1}$$
$$> p^M(\theta_i)$$
$$q^f(\theta_i) \equiv q^f(p^f(\theta_i),\theta_i) = q^M(\theta_i),$$

where $q^M(\theta_i)$ and $p^M(\theta_i)$ denote the profit-maximizing monopoly quantity and price for $\theta = \theta_i$. The additional demand due to expected damages exactly compensates for the deterrent effects of damages. Thus, if a multiple damages remedy does not totally deter price-fixing (i.e., $\beta t < 1$), the cartel chooses the monopoly output and charges a price that is greater than the monopoly price. Moreover, it is straightforward to verify that consumers' surplus $cs^f(\theta_i)$ and cartel profits $\pi^f(\theta_i)$ under private antitrust are equal to their respective levels, $cs^M(\theta_i)$ and $\pi^M(\theta_i)$, in the absence of antitrust. The gain to consumers from expected damage recovery is exactly offset by the loss in consumers' surplus due to the higher price. One can show that if the probability of conviction β depends on the markup, as in Baker (1988) and Salant (1987), the neutrality result is preserved.

A number of institutional factors can eliminate the neutrality of treble damages.

[7] Baker (1988) recognizes that private information may be inferred from the cartel's market actions but exogenously specifies buyer expectations to reflect either full information or a windfall damage recovery.

[8] We assume that the profit function $(p - \theta_i)q^f(p,\theta_i)(1-\beta t)$ is well-behaved: twice continuously differentiable in p with a unique interior local and global maximum on (θ_i,∞).

Limited liability can cause the cartel price to fall below the monopoly level. Spulber (1989) shows that for the intermediate liability levels, some moderation of markups may occur. However, relatively high liability levels do not affect the equilibrium and relatively low liability levels are simply deducted from cartel profits with no equilibrium effects. In a dynamic setting, Spiller (1986) shows that with limited liability, consumers may postpone their suits perhaps indefinitely or until damages rise to equal the firm's liability. Spulber (1989) also considers the case of fixed costs of suit. Fixed legal costs can cause the cartel to set a price $\tilde{p}$ such that the expected damage payment exactly equals the expected cost of suit k,

$$\beta t(\tilde{p}-\theta)q(\tilde{p})=k,$$

where $q(p)$ is market demand. Thus, markups are positive, but the consumer has no incentive to sue. For example, under treble damages, the cartel's profits equal one-third of consumer legal costs divided by the likelihood of conviction; that is, $\pi = k/3\beta$. This holds even if the costs of a successful suit are borne by the defendant (by Section 4 of the Clayton Act), because we can let $k=(1-\beta)k'$ where k' is the actual cost of suit.

Suppose that damages are calculated on the basis of a reference price $\bar{p}$ in excess of marginal cost (i.e., $(p-\bar{p})q$). The cartel is thus allowed a maximum positive markup. Then cartel actions can be affected by a variety of institutional constraints. "Decoupling" of damage payments and penalties for price fixing, as advocated by William Schwartz (1980), A. Mitchell Polinsky (1986), and others, moderates cartel markups. Under decoupling, the government imposes a fine on price fixers that drives a wedge between damage multiples for the cartel and for consumers, $t^c \neq t^f$. Similarly, if consumers and firms face legal costs that are contingent and proportional to damages, and these legal costs are not equal, treble damages are no longer neutral. Finally, suppose that the consumer and firm have differing estimates of the likelihood of conviction, $\beta^c \neq \beta^f$. Then, the damage remedy moderates markups and is no longer neutral.[9] In each of these cases, setting the reference price equal to marginal cost restores the neutrality result.

III. Sequential Equilibrium Under Asymmetric Information

Consumers who have imperfect information about the cartel's production costs cannot accurately assess the extent of the violation and the size of damages. In this section, we examine the impact of asymmetric information using a sequential equilibrium model of the interaction of a representative consumer and the cartel. The analysis indicates that asymmetric information introduces a friction that can eliminate the neutrality of a multiple damages remedy.

The game between the cartel and consumers is defined as follows. A cartel with cost $\theta_i \in \{\theta_1,\dots,\theta_m\}$ chooses a pricing strategy $p=\rho(\theta_i)$.[10] The representative consumer observes the price p set by the cartel and makes an inference, represented by a posterior probability $\gamma_i(p)$, that the cartel's costs are equal to θ_i. Let $\gamma(p)=(\gamma_1(p),\dots,\gamma_m(p))$ represent the posterior probability distribution across cartel types.[11] The consumer makes his purchase decision $q(p)$ and his decision to sue to maximize expected utility, given the observed price p and the posterior probability distribution $\gamma(p)$ on the cartel's costs. To keep the focus entirely on the effects of asymmetric information, assume zero legal costs for both the cartel and the consumer. Because legal costs are zero, the consumer will sue as long the observed price p exceeds the lowest possible cost θ_1 of the cartel.

A *sequential equilibrium*[12] of the game consists of strategies and beliefs $\{\rho^*(\theta_i),$

[9] This is shown in Spulber (1989). Baker (1988) also examines the case of divergent estimates of the likelihood of conviction.

[10] Attention is restricted to pure strategies for ease of presentation although it can be shown that the cartel will not employ randomized strategies.

[11] Thus $0 \leq \gamma_i(p) \leq 1$, $i=1,\dots,m$ and $\Sigma_{i=1}^{m}\gamma_i(p)=1$.

[12] See David Kreps and Robert Wilson (1982) for a general definition of sequential equilibrium.

$q^*(p), \gamma^*(p)\}$ such that the following conditions hold.

(a) For each $\theta_i \in \{\theta_1, \ldots, \theta_m\}$, the cartel's pricing strategy $\rho^*(\theta_i)$ is a profit-maximizing best response to the consumer's purchasing strategy. That is, $\rho^*(\theta_i)$ maximizes

$$(p-\theta_i)q^*(p)(1-\beta t)$$

for every $\theta_i \in \{\theta_1, \ldots, \theta_m\}$.

(b) For any $p \geq 0$ the consumer's purchasing strategy $q^*(p)$, given beliefs $\gamma^*(p)$ and the observed market price p, maximizes net expected benefits,

$$V(q) - pq + \beta t(p - \bar{\theta}^*(p))qI_{[\theta_1, \infty)}(p),$$

where $\bar{\theta}^*(p) \equiv \Sigma_{i=1}^m \theta_i \gamma_i^*(p)$ is the consumer's posterior expectation of marginal cost. This condition implies that for $p > \theta_1$,

$$q^*(p) = q^f(p, \bar{\theta}^*(p)).$$

(c) For prices p on the equilibrium path—that is, $\rho^{*-1}(p) \neq \varnothing$—the consumer's beliefs $\gamma^*(p)$ are consistent with Bayes' rule and the cartel's equilibrium strategy $q^*(p)$. This implies

$$\gamma_i^*(p) = \begin{cases} \dfrac{\gamma_i}{\sum_{\{j:\rho^*(\theta_j)=p\}} \gamma_j} \\ \quad \text{if } \rho^*(\theta_i) = p \\ 0 \\ \quad \text{if } \rho^*(\theta_i) \neq p. \end{cases}$$

A price p is said to be "off the equilibrium path" if it would not be chosen by a cartel following the equilibrium strategy $\rho^*(\cdot)$. That is, a price p such that $\rho^{*-1}(p) = \varnothing$ is off the equilibrium path. The definition of a sequential equilibrium requires specification of the consumer's beliefs in response to off the equilibrium path prices. This is required to verify that the cartel will not have an incentive to deviate from its equilibrium strategy. In many sequential equilibrium models, the degrees of freedom afforded by off the equilibrium path beliefs results in multiple sequential equilibria. In this analysis, we restrict off the equilibrium path beliefs to satisfy the universal divinity criterion due to Banks and Sobel (1987). Following Banks (1988), universal divinity is defined as follows.

Let $\rho^*(\theta_i)$ and $q^*(p)$ be equilibrium strategies, and let p be a price such that $\rho^{*-1}(p) = \varnothing$. Define the function $h(\theta_i, p)$ as the level of consumer demand that makes a cartel with cost θ_i indifferent between following its equilibrium strategy $\rho^*(\theta_i)$ and deviating from that strategy by charging price p; that is, $h(\theta_i, p)$ is given by

$$(3) \quad (p-\theta_i)h(\theta_i, p)(1-\beta t)$$
$$= (\rho^*(\theta_i) - \theta_i)q^*(\rho^*(\theta_i))(1-\beta t),$$

for each $\theta_i \in \{\theta_1, \ldots, \theta_m\}$.[13] *Universal divinity* requires that consumer beliefs place positive probability only on those types that are most likely to deviate from the optimal strategy. For any p, this requires finding the *lowest* critical output level $h(\theta_i, p)$ across all possible θ_i. This is because a cartel anticipating a demand response larger than $h(\theta_i, p)$ will also choose to deviate. Thus, given the universal divinity criterion, out-of-equilibrium beliefs place positive probability on those types θ' such that[14]

$$\theta' \in \arg \min_{\theta \in \{\theta_1, \ldots, \theta_m\}} h(\theta, p).$$

We develop the characterization of the sequential equilibrium in a sequence of lemmas, culminating in a characterization

[13] Because the cartel's equilibrium profit must be positive, condition (3) defines $h(\theta_i, p)$ only for $p > \theta_i$. For $p \leq \theta_i$, we define $h(\theta_i, p) = \infty$.

[14] Note that by defining $h(\theta_i, p) = \infty$ for $\theta_i > p$, we ensure that the consumer places zero probability weight on those types θ_i for which $\theta_i > p$. This is as it should be. A firm with cost type θ_i would never wish to deviate from its equilibrium price to a price p below its marginal cost, no matter what the anticipated demand response.

of the equilibrium in Propositions 1, 2, and 3. A discussion of the intuition underlying the equilibrium follows Proposition 3.

Let $\pi^*(\theta_i)$ denote the equilibrium profits of a cartel with cost θ_i:

$$(4)\quad \pi^*(\theta_i)=(\rho^*(\theta_i)-\theta_i)q^*(\rho^*(\theta_i))\times(1-\beta t), i\in\{1,\dots,m\}.$$

For the pricing strategy $\rho^*(\theta_i)$ to be a sequential equilibrium strategy, it must be *incentive compatible.* That is, a cartel with cost θ_i must prefer the price $\rho^*(\theta_i)$ to the price $\rho^*(\theta_j)$ that a cartel without θ_j would choose in equilibrium. Thus,

$$(5)\quad \pi^*(\theta_i)\geq(\rho^*(\theta_j)-\theta_i)q^*(\rho^*(\theta_j))\times(1-\beta t), i,j\in\{1,\dots,m\}, i\neq j.$$

Incentive compatibility has the following important implications.

LEMMA 1: *The cartel's equilibrium pricing strategy is nondecreasing in θ_i; that is,*

$$\rho^*(\theta_1)\leq\rho^*(\theta_2)\leq\cdots\leq\rho^*(\theta_m).$$

Along the equilibrium path, the equilibrium quantity demanded is nonincreasing in θ_i; that is,

$$q^*(\rho^*(\theta_1))\geq q^*(\rho^*(\theta_2))\geq\cdots\geq q^*(\rho^*(\theta_m)).$$

Finally, the cartel's equilibrium profit is decreasing in θ_i; that is,

$$\pi^*(\theta_1)>\pi^*(\theta_2)>\cdots>\pi^*(\theta_m).$$

The proofs of this and all subsequent lemmas and propositions appear in the Appendix.

The next step in the analysis is to use the universal divinity criterion to pin down equilibrium posterior beliefs.

LEMMA 2: *Let $\{\rho^*(\cdot),q^*(\cdot),\gamma^*(\cdot)\}$ be a sequential equilibrium satisfying universal divinity. Suppose that the equilibrium is fully separating; that is, $\rho^*(\theta_1)<\rho^*(\theta_2)<\cdots<\rho^*(\theta_m)$. Then equilibrium posterior beliefs are given by*[15]

$$\begin{aligned}
&\gamma_1^*(p)=1 && p\in(\theta_1,\bar{p}_1]\\
&\gamma_i^*(p)=1 && p\in(\bar{p}_{i-1},\bar{p}_i],\\
& && i\in\{2,\dots,m-1\}\\
&\gamma_m^*(p)=1 && p>\bar{p}_{m-1},
\end{aligned}$$

where

$$\bar{p}_i\equiv\theta_i+\frac{\pi^*(\theta_i)(\theta_{i+1}-\theta_i)}{\pi^*(\theta_i)-\pi^*(\theta_{i+1})},\quad i\in\{1,\dots,m-1\},$$

$$\bar{p}_i\leq\bar{p}_{i+1};$$

and

$$\bar{p}_i\in[\rho^*(\theta_i),\rho^*(\theta_{i+1})]\ \textit{for}\ i\in\{1\dots m-1\}.$$

This lemma says that for off-the-equilibrium path prices within a neighborhood $(\bar{p}_{i-1},\bar{p}_i)$ of the equilibrium price $\rho^*(\theta_i)$, a consumer infers that the cartel's cost is θ_i, with probability one. For prices less than (greater than) the smallest (largest) possible equilibrium price, the consumer infers that the cartel has the lowest (highest) θ.

The possibility that the sequential equilibrium involves pooling is ruled out by the universal divinity restriction on beliefs.

LEMMA 3: *Let $\{\rho^*(\cdot),q^*(\cdot),\gamma^*(\cdot)\}$ be a sequential equilibrium satisfying universal divinity. The equilibrium must be fully separating; that is,*

$$\rho^*(\theta_1)<\rho^*(\theta_2)<\cdots<\rho^*(\theta_m)$$

$$q^*(\rho^*(\theta_1))>q^*(\rho^*(\theta_2))>\cdots>q^*(\rho^*(\theta_m)).$$

[15] For $p\leq\theta_1$ consumer beliefs are not relevant because consumers will not sue for price fixing.

The universally divine sequential equilibrium is now characterized in Proposition 1.

PROPOSITION 1: *There is a unique sequential equilibrium satisfying universal divinity. That equilibrium has the following features.*

(*a*) *A cartel with the highest possible cost θ_m produces the full-information (i.e., monopoly) quantity, charges the full-information price, and earns its full-information (i.e., monopoly) profits,*

$$q^*(\rho^*(\theta_m)) = q^f(\theta_m) = q^M(\theta_m),$$

$$\rho^*(\theta_m) = p^f(\theta_m) > p^M(\theta_m),$$

$$\pi^*(\theta_m) = \pi^f(\theta_m) = \pi^M(\theta_m).$$

(*b*) *The equilibrium price $\rho^*(\theta_i)$ for a cartel with cost $\theta_i, i \in \{1,\dots,m-1\}$, solves the following constrained optimization problem*

$$\max_p (p-\theta_i)q^f(p,\theta_i)(1-\beta t)$$

$$\text{(6)}\quad \textit{subject to}: (p-\theta_{i+1})q^f(p,\theta_i)(1-\beta t)$$

$$\leq \pi^*(\theta_{i+1}).$$

(*c*) *Equilibrium consumer demand is given by*

$$q^*(p) = \begin{cases} Q(p) & p \in [0,\theta_1] \\ q^f(p,\theta_1) & p \in (\theta_1, \bar{p}_1] \\ q^f(p,\theta_i) & p \in (\bar{p}_{i-1}, \bar{p}_i], \\ & i \in \{2,\dots,m-1\} \\ q^f(p,\theta_m) & p > \bar{p}_{m-1}. \end{cases}$$

A multiple damages remedy is only neutral in general when the cartel has the highest possible cost, as shown by part (a) of Proposition 1. A cartel with the highest possible cost chooses the monopoly output and earns monopoly profit.[16] The highest-cost cartel's price exceeds the monopoly price by just the amount needed to compensate for the damage award, as in the full-information case. A cartel with any lower cost behaves *as if* it were solving a constrained optimization problem, given as equation (6) in part (b). The constraint arises as a result of consumer expectations. Clearly, a cartel would like to choose the profit-maximizing price. However, if the cartel's price is set too high, consumers will incorrectly infer that the cartel's costs are higher than they actually are. This will reduce consumer demand because consumers would then anticipate a lower damage award. The reduction in demand would result in lower cartel profits. Therefore, the cartel's pricing decision is constrained.

The effects of the constraint imposed on the cartel by consumer expectations are reflected in part (c) of Proposition 1. It is shown that if the cartel's price falls below a critical value, consumers' expectations of cartel costs are revised downward and consumers adjust their demand accordingly.[17] This yields a modified market demand schedule, $q^*(p)$, that cuts across the family of full-information demand schedules, as shown in Figure 1.

The unique sequential equilibrium obtained in Proposition 1 has a significant economic implication. Private antitrust enforcement effectively alters the market equilibrium and improves social welfare whenever consumer expectations constrain cartel pricing behavior.

[16] The observation that the equilibrium outcome at one end of the range of agent types corresponds to the full-information outcome is relatively standard in signaling models. See, particularly, George Mailath (1987) on separating equilibria in a continuous-types signaling game. It is expected that our result will carry over to the case of a continuum of types, as in Mailath. Then in any separating equilibrium, all lower cost types will price below their full-information price, thus generating welfare gains.

[17] If constraint (6) in the optimization problem in Proposition 1 is binding for a given i, then $\bar{p}_i = \rho^*(\theta_i)$. This implies that an increase in price by a cartel with cost θ_i above its equilibrium price will induce consumers to infer that its cost is greater than θ_i.

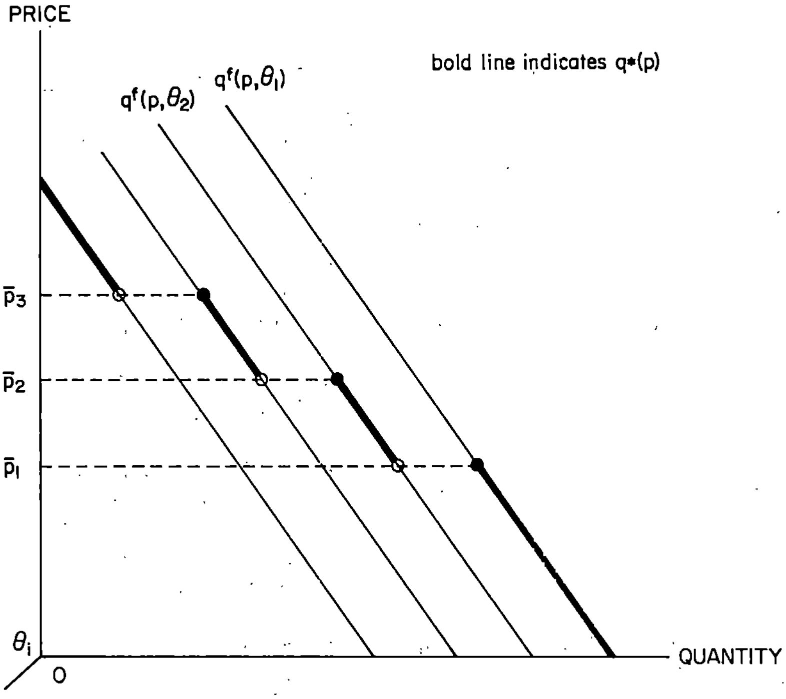


Figure 1. Equilibrium Demand Curve $q^*(p)$

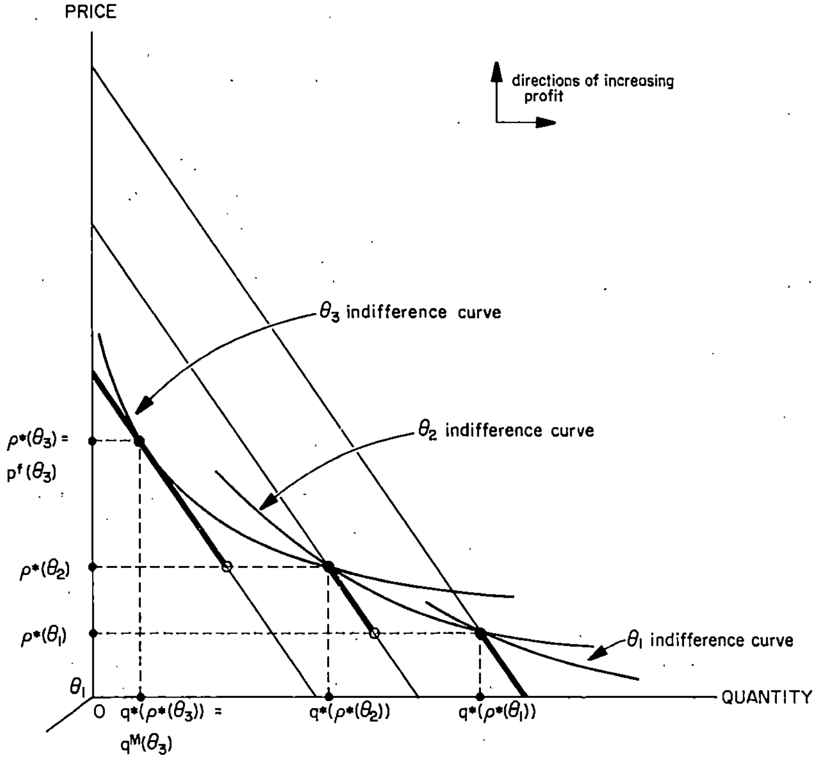


Figure 2. Sequential Equilibrium for the Three-Type Case

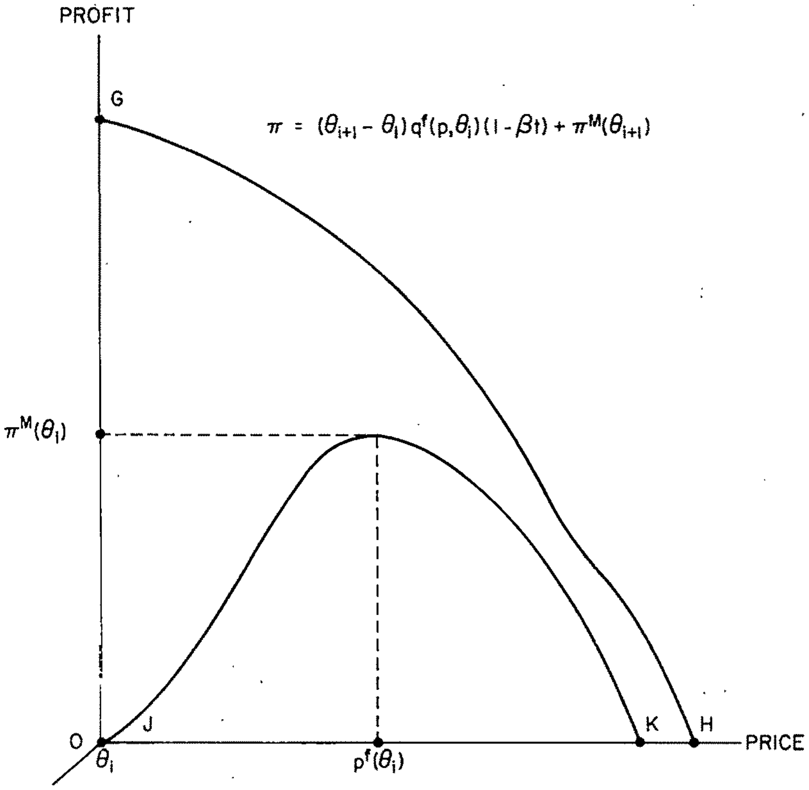

PROFIT
G
$\pi = (\theta_{i+1} - \theta_i) q^f(p,\theta_i)(1 - \beta t) + \pi^M(\theta_{i+1})$
$\pi^M(\theta_i)$
O
J
θ_i
$p^f(\theta_i)$
K
H
PRICE

FIGURE 3. THE SEQUENTIAL EQUILIBRIUM IS WELFARE NEUTRAL WHEN CARTEL'S COST IS θ_i

PROPOSITION 2: *Private antitrust enforcement under a multiple damages remedy lowers the equilibrium price, increases output, and raises social welfare whenever the consumer expectations constraint* (6) *is binding. That is, if* (6) *is binding for any given i, then* $p^*(\theta_i) < p^f(\theta_i)$ *and* $q^*(p^*(\theta_i)) > q^f(\theta_i) = q^M(\theta_i)$.

The sequential equilibrium for a three-type example is illustrated in Figure 2. The downward sloping convex curves are profit indifference curves for the cartel in quantity-price space. The downward sloping lines are the full-information demand functions, $q^f(p,\theta_i)$. The bold line indicates the equilibrium consumer demand function, $q^*(p)$. The figure shows the case in which constraint (6) is binding for $i = 1$ and 2. In this case, the equilibrium price for a cartel with either of these cost levels is less than the full-information price.

Proposition 2 establishes a condition for the nonneutrality of private antitrust enforcement. This is useful in establishing a critical value for the damage multiple that yields effective private enforcement. This is the subject of the next section.

IV. Welfare-Improving Antitrust Enforcement

The situations in which private antitrust enforcement enhances social welfare can be illustrated with the aid of Figures 3 and 4. Figure 3 depicts the constrained optimization problem for a cartel with costs θ_i. The hill JK shows attainable price-profit combinations for the cartel. Suppose private enforcement is welfare neutral for the next highest cost level θ_{i+1}. The locus GH then shows the boundary of the set of points satisfying (6). The full-information solution is optimal if and only if the locus GH lies

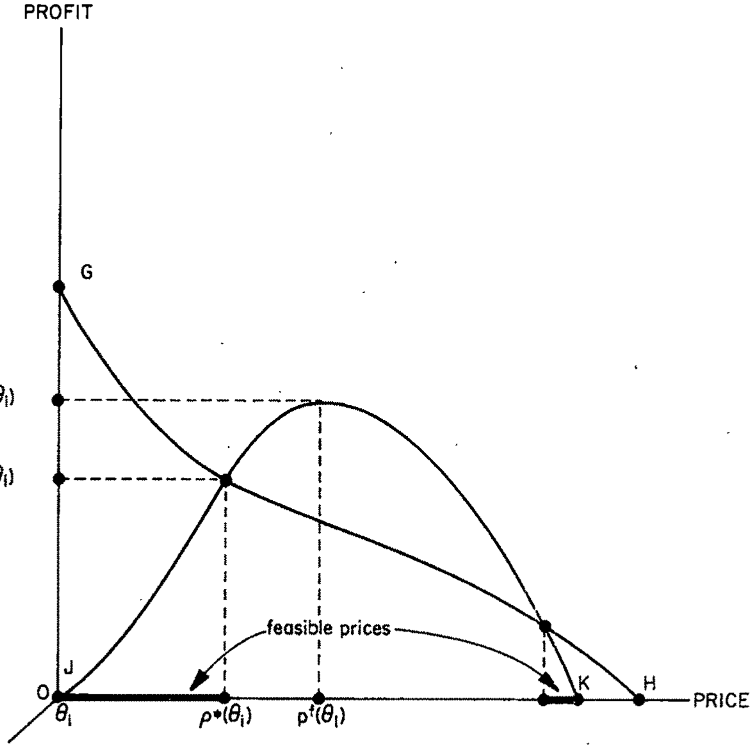


FIGURE 4. THE SEQUENTIAL EQUILIBRIUM IS WELFARE ENHANCING WHEN CARTEL'S COST IS θ_i

above the profit hill at the full-information price $p^f(\theta_i)$, as shown in Figure 3. Algebraically this is equivalent to[18]

$$\pi^M(\theta_i) \le \pi^M(\theta_{i+1}) + (\theta_{i+1}-\theta_i)q^M(\theta_i)(1-\beta t)$$

or

$$(7)\quad \beta t \le \phi_i \equiv 1-(\pi^M(\theta_i)-\pi^M(\theta_{i+1})) /((\theta_{i+1}-\theta_i)q^M(\theta_i)) \in (0,1).$$

If condition (7) is not satisfied, then the situation is as shown in Figure 4, and it follows that the sequential equilibrium price $\rho^*(\theta_i)$ is strictly less than the full-information price $p^f(\theta_i)$.

The condition in (7) enables us to prove the following result.

PROPOSITION 3.

(a) *If the expected damage multiple βt is less than or equal to a critical value $\phi^* \equiv min\{\phi_1,\dots,\phi_{m-1}\}$, then for all $i \in \{1,\dots,m\}$ $\rho^*(\theta_i) = p^f(\theta_i)$, and $q^*(\rho^*(\theta_i)) = q^M(\theta_i)$. That is, private antitrust enforcement is welfare neutral for any possible realization of cartel cost.*

(b) *If the expected damage multiple βt exceeds the critical value ϕ^*—that is, $\beta t \in (\phi^*,1)$—then there exists at least one $i \in \{1,\dots,m-1\}$ such that $\rho^*(\theta_i) < p^f(\theta_i)$ and $q^*(\rho^*(\theta_i)) > q^M(\theta_i)$. That is, private antitrust enforcement results in welfare improvements for some realizations of cartel cost.*

[18] The proof that $\phi_i > 0$ is as follows. From the envelope theorem, $\pi^M(\theta)$ is decreasing and convex in θ with $d\pi^M(\theta)/d\theta = -q^M(\theta)$. By convexity, $q^M(\theta_i) > (\pi^M(\theta_i) - \pi^M(\theta_{i+1}))/(\theta_{i+1}-\theta_i)$.

Proposition 3 indicates that a relatively high damage multiple, such as treble damages, is more likely to lead to welfare improvements. Detrebling antitrust damages may push the expected multiple below the critical level, ϕ^*.

Given a sufficiently high conviction probability or damage multiple, private antitrust enforcement is unlikely to be welfare neutral under asymmetric information. Moreover, as θ_{i+1} approaches θ_i for each i, the critical value ϕ^* approaches zero and private enforcement results in marginal deterrence for nearly all values of β and t. Thus, unless significant differences exist between possible cartel costs, private antitrust enforcement is unlikely to be neutral.

The intuition for why asymmetric information can upset the neutrality results has to do with the relationship between the cartel's price and consumers' expectations of the magnitude of the antitrust violation. When consumers do not know the cartel's cost, they do not know with certainty the damages they will recover if the cartel is convicted of a violation. However, they know that for any given market price, damages will be greater the lower is the cartel's cost. In equilibrium, a cartel with a lower cost charges a lower market price, and consumers anticipate this. As the cartel lowers its market price, the consumer's net price—price minus the expected damage award—falls.[19] Due to the informational effect, the net price falls by a greater amount than it would under full information.[20] Thus, under asymmetric information, a given decrease in market price generates more additional demand than it would under full information. In effect, asymmetric information makes the market demand curve more elastic, which leads the cartel to set a lower price than it would under full information. This effect can be seen most clearly for the case in which the possible realizations of the cartel's costs are very close together. In this case, the equilibrium demand curve approaches the smooth curve shown in Figure 5.

An example can be given to illustrate the situation where antitrust is always effective for low cost realizations. For the case of linear demand, $P(q) = a - bq$, and a discrete uniform distribution of types with $\theta_{i+1} - \theta_i = \xi$ for $i \in \{1,\dots,m-1\}$ and $\beta t > \frac{1}{2}\xi(a-\theta_1)^{-1}$, there exists a cost type $\theta_j > \theta_1$ such that private antitrust enforcement results in a welfare improvement for all cost realizations $\theta_i < \theta_j$ and results in welfare neutrality for all cost realizations $\theta_i \geq \theta_j$.[21] That is,

$$\rho^*(\theta_i) < p^f(\theta_i) \quad \text{and}$$
$$q^*(\rho^*(\theta_i)) > q^M(\theta_i) \text{ for } i \in \{1,\dots,j-1\},$$
$$\rho^*(\theta_i) = p^f(\theta_i) \quad \text{and}$$
$$q^*(\rho^*(\theta_i)) = q^M(\theta_i) \text{ for } i \in \{j,\dots,m\}.$$

The expected damage multiple βt has two effects on the market equilibrium. First, a higher expected damage multiple can raise the likelihood that antitrust policy will be welfare enhancing by increasing the set of cost types for which nonneutrality holds, as in the preceding example. Second, as shown in the next proposition, given nonneutrality, a higher expected damage multiple induces the cartel to moderate its markup.

PROPOSITION 4: *Suppose $q^*(\theta_i) \equiv q^*(\rho^*(\theta_i)) < q^M(\theta_i)$; that is, the equilibrium results in a welfare improvement when the*

[19] We would like to thank Jonathan Baker for suggesting this verbal exposition.

[20] The buyer's net price, p^{NET}, is given by

$$p^{NET} = p - \beta t(p - \theta^*(p))$$
$$= (1-\beta t)p + \beta t\theta^*(p),$$

where p is the market price. Under full information, $\Delta p^{NET}/\Delta p = 1 - \beta t$, while under asymmetric information,

$$\Delta p^{NET}/\Delta p = (1-\beta t) + \beta t \Delta\theta^*/\Delta p.$$

From Lemma 2, $\Delta\theta^*/\Delta p$, the change in consumers' expectations with respect to price, is positive. Thus $\Delta p^{NET}/\Delta p$ is larger under asymmetric information.

[21] See Appendix for a proof.

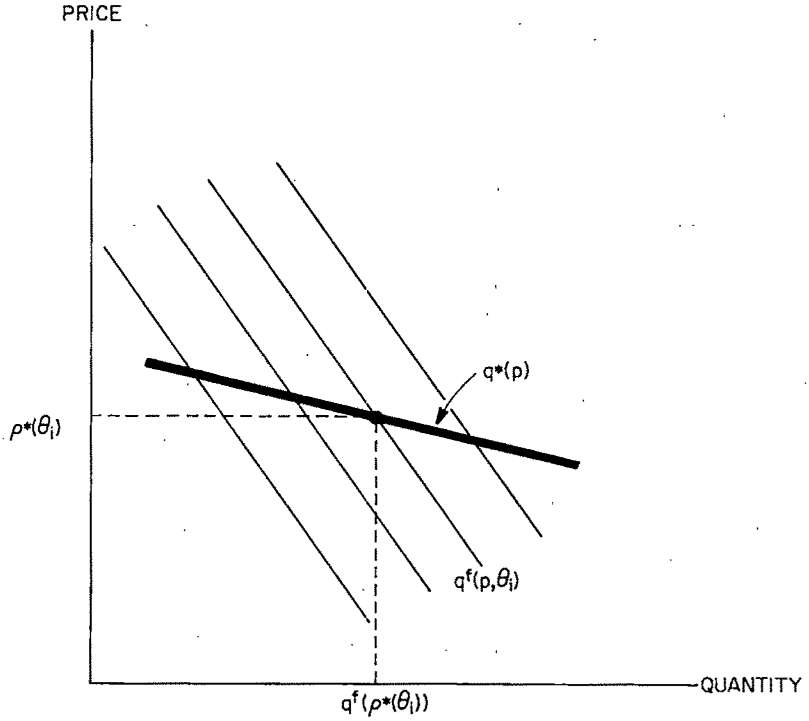


FIGURE 5. SEQUENTIAL EQUILIBRIUM DEMAND CURVE AS $\theta_i \rightarrow \theta_{i+1}$ FOR ALL i

cartel's cost is θ_i. An increase in the conviction probability β or the damage multiple t increases the equilibrium quantity $q^*(\theta_i)$, decreases the equilibrium price $p^*(\theta_i)$, and thus increases social welfare.

Proposition 4 emphasizes the marginal deterrent effect of antitrust under asymmetric information. This implies that policy makers can influence both cartel pricing behavior and consumer enforcement decisions by the expected damage multiple. In particular, policymakers can increase the damage multiple t or raise the likelihood of conviction β through such procedural changes as rules of evidence or through public enforcement efforts.

The basic model can be generalized to incorporate various institutional features. It is expected that the nonneutrality of private antitrust enforcement would be robust to these changes. Legal costs would be expected to moderate markups if they lie in a critical range, as in the full-information setting.[22] A reference price for calculating damages that lies above marginal cost would strongly reinforce our nonneutrality result. If the consumer and firm differ in their estimates of the likelihood of conviction, this would further enhance the effects of private enforcement. Decoupling of damage payments and penalties would also be expected to enhance the effectiveness of private enforcement under asymmetric information, at least in the presence of a reference price. An interesting extension of

[22] An interesting aspect of positive legal costs for consumers and firms is that it may lead to a selection bias in terms of the cost types that are actually sued. If consumers have positive fixed legal costs (as opposed to legal costs that are proportional to the damage award) it will not be worthwhile for them to sue if the expected markup is sufficiently low. This suggests that, in equilibrium, there will be some cartel types that are not sued. Nevertheless, private antitrust enforcement may still be effective even though a cartel is not sued since the cartel may lower its markup to avoid incurring the legal costs of suit.

the present model would be to allow for pretrial settlements. This might affect the deterrence properties of private enforcement. Also, the presence of asymmetric information would complicate the bargaining over the out-of-court settlement.

V. Conclusion

Private antitrust enforcement with a multiple damages remedy can enhance social welfare if consumers lack complete information about the cartel's production costs. At the sequential equilibrium, the cartel price conveys cost information to consumers. On the basis of their updated expectations, consumers make demand decisions and estimate potential damage awards. This yields a modified market demand schedule for the cartel. At the resulting market equilibrium, the cartel's markup is moderated, leading to increased output and social welfare.

The sequential equilibrium model of private antitrust enforcement indicates that asymmetric information can increase social welfare. This is in contrast to the full-information models of Baker (1988) and Salant (1987) in which neutrality arises. In many applied game-theoretic models, informational asymmetries create distortions in decision making that reduce social welfare. The gains from information asymmetry obtained here are the result of mitigation of monopoly price fixing. Under full information, consumer demand and cartel markups perfectly adjust for antitrust damages. However, it is reasonable to expect that consumers can only imperfectly estimate potential damages. In fact, the presence of asymmetric information alleviates the "perverse effect" of increased consumer purchases to alter the damage award. Collusive firms must then moderate price increases in equilibrium to avoid raising consumer estimates of costs and the resulting lowering of demand.

The analysis suggests a connection between public and private antitrust enforcement. Private antitrust enforcement efforts generally occur after successful public enforcement. Elzinga and Breit (1976, p. 142) attribute this to the fact that "the private sector still does not have the 'fishing rights' available to the government to investigate possible antitrust violations prior to an actual case." Suppose that the likelihood of a private plaintiff winning a case, β_1, rises to $\beta_2 > \beta_1$ following a successful public antitrust suit. Proposition 3 obtained a critical minimum level ϕ^* of the expected damage multiple for which private antitrust enforcement can enhance welfare. Suppose that $\beta_2 t > \phi^* > \beta_1 t$. Then, public enforcement effort, by raising the expected damage multiple in price-fixing cases, can enhance the effectiveness of private enforcement.

APPENDIX

PROOF OF LEMMA 1: The incentive compatibility condition (5) implies, for $\theta_j > \theta_i$,

$$(\theta_j - \theta_i) q^*\big(\rho^*(\theta_j)\big)(1-\beta t) \le \pi^*(\theta_i) - \pi^*(\theta_j)$$

$$\le (\theta_j - \theta_i) q^*\big(\rho^*(\theta_i)\big)(1-\beta t).$$

This implies $q^*(\rho^*(\theta_i)) \ge q^*(\rho^*(\theta_j))$ and $\pi^*(\theta_i) > \pi^*(\theta_j)$.

Now, suppose $\theta_i > \theta_j$. We have just established that $q^*(\rho^*(\theta_i)) \le q^*(\rho^*(\theta_j))$. A necessary condition for (5) to hold is $\rho^*(\theta_i) \ge \rho^*(\theta_j)$. □

PROOF OF LEMMA 2: As a first step, we prove that $\bar{p}_i \in [\rho^*(\theta_i), \rho^*(\theta_{i+1})]$ for any $i \in \{1, \dots, m-1\}$. The incentive compatibility condition (5) implies

$$\text{(A1)} \quad \pi^*(\theta_i) - \pi^*(\theta_{i+1}) \le (\theta_{i+1} - \theta_i) q^*\big(\rho^*(\theta_i)\big)(1-\beta t).$$

Thus

$$\bar{p}_i \ge \theta_i + \frac{\pi^*(\theta_i)}{q^*\big(\rho^*(\theta_i)\big)(1-\beta t)} = \rho^*(\theta_i).$$

A symmetric argument establishes $\bar{p}_i \ge \rho^*(\theta_{i+1})$. Thus, in general (that is, whether the equilibrium is fully separating or not), $\bar{p}_{i-1} \le \rho^*(\theta_i) \le \bar{p}_i \le \rho^*(\theta_{i+1}) \le \bar{p}_{i+1}$. Now, in a fully separating equilibrium, $\rho^*(\theta)$ is strictly increasing in θ, so it follows immediately that $\bar{p}_i < \bar{p}_{i-1}$. Moreover, because $\rho^*(\theta_i) \in [\bar{p}_{i-1}, \bar{p}_i]$, for prices that are on the equilibrium path, the beliefs specified in the lemma are consistent with Bayes' rule.

Now, as a next step, fix an arbitrary $i \in \{2, \dots, m-1\}$ and $p \in (\bar{p}_{i-1}, \bar{p}_i]$. We will now show that

$$\text{(A2)} \qquad \theta_i = \arg \min_{\theta \in \{\theta_1, \dots, \theta_m\}} h(\theta, p).$$

To prove (A2) note, first, that for any k

$$\text{(A3)}\quad h(\theta_k,p)-h(\theta_{k+1},p)=\frac{\pi^*(\theta_k)}{p-\theta_k}-\frac{\pi^*(\theta_{k+1})}{p-\theta_{k+1}}$$

$$=\frac{\pi^*(\theta_k)-\pi^*(\theta_{k+1})}{(p-\theta_k)(p-\theta_{k+1})}[p-\bar{p}_k].$$

We are given that $p\le\bar{p}_i$, so it is indeed the case that[23]

$$h(\theta_i,p)\le h(\theta_{i+1},p).$$

Now, recall $\bar{p}_i\le\bar{p}_{i+1}$. Employing (A3) again (now with $k=i+1$) yields

$$h(\theta_{i+1},p)\le h(\theta_{i+2},p).$$

Iterating upward in this fashion for $i+2,\ i+3,\dots,m$ yields the result that

$$\text{(A4)}\quad h(\theta_i,p)\le h(\theta_j,p)\ \text{for}\ p\le\bar{p}_i,$$

$$j\in\{i+1,\dots,m\}.$$

Now, we are also given that $p>\bar{p}_{i-1}$. Again using (A3) (with $k=i-1$), it follows that

$$h(\theta_{i-1},p)\ge h(\theta_i,p).$$

Moreover, from our result above, $p>\bar{p}_{i-1}\ge\bar{p}_{i-2}$. Once again using (A3) (with $k=i-2$) yields

$$h(\theta_{i-2},p)>h(\theta_{i-1},p).$$

Iterating downward in this fashion for $i-3,\dots,1$ yields the result that

$$\text{(A5)}\quad h(\theta_i,p)\le h(\theta_j,p)\ \text{for}\ p>\bar{p}_{i-1},$$

$$j\in\{i-1,\dots,m\}.$$

For $p\in(\bar{p}_{i-1},\bar{p}_i]$, conditions (A4) and (A5) simultaneously hold, so

$$\theta_i=\arg\min_{\theta\in\{\theta_1,\dots,\theta_m\}}h(\theta,p).$$

The definition of universal divinity thus requires beliefs given by

$$\gamma_i^*(p)=1\ \text{for}\ p\in(\bar{p}_{i-1},\bar{p}_i],i\in\{2,\dots,m-1\}.$$

For $p\in(\theta_1,\bar{p}_1]$, logic that is identical to that just employed leads to the result that

$$h(\theta_1,p)\le h(\theta_j,p),\quad\text{for}$$

$$j\in\{2,\dots,m\},\ p\in(\theta_1,p_1].$$

Universal divinity thus requires $\gamma_1^*(p)=1$ for $p\le\bar{p}_1$. Similarly, for $p>\bar{p}_{m-1}$ condition (A3) can be shown to imply

$$h(\theta_m,p)\le h(\theta_j,p),\quad\text{for}\ j\in\{1,\dots,m-1\},$$

$$p>\bar{p}_{m-1}.$$

Thus, universal divinity requires $\gamma_m^*(p)=1$ for $p>\bar{p}_{m-1}$. □

PROOF OF LEMMA 3: Suppose, to the contrary, that the equilibrium involves pooling. Let $T\subseteq\{\theta_1,\dots,\theta_m\}$ be a pooling set whose smallest element is θ_T. For $\theta\in T$,

$$\text{(A6)}\quad \pi^*(\theta)=(\rho^*(\theta_T)-\theta)q^*(\rho^*(\theta_T))$$

$$=(\rho^*(\theta_T)-\theta)q^f(\rho^*(\theta_T),\bar{\theta}(T)),$$

where

$$\text{(A7)}\quad \bar{\theta}(T)\equiv\left[\sum_{\{j:\theta_j\in T\}}\gamma_j\theta_j\right]\Big/\left[\sum_{\{j:\theta_j\in T\}}\gamma_j\right]>\theta_T.$$

That is, $\bar{\theta}(T)$ is the conditional expectation of the cartel's cost, given that θ is in the set T. Condition (A7) follows because equilibrium beliefs must satisfy Bayes rule.

Now consider a price $p(\varepsilon)=\rho^*(\theta_T)-\varepsilon$ where $\varepsilon>0$. Using the logic in the proof of the previous lemma, one can show that

$$\text{(A8)}\quad \theta_T\ge\arg\min_{\theta\in\{\theta_1,\dots,\theta_m\}}h(\theta,p(\varepsilon)).$$

By (A8), universal divinity requires that for p slightly less than $\rho^*(\theta_T)$, consumers place positive probability weight only on cost realizations of θ_T or smaller. This implies $\bar{\theta}^*(p(\varepsilon))\le\theta_T<\bar{\theta}(T)$. Thus

$$\text{(A9)}\quad \lim_{\varepsilon\downarrow 0}(p(\varepsilon)-\theta)q^*(p(\varepsilon))$$

$$=(\rho^*(\theta_T)-\theta)q^f\Big(\rho^*(\theta_T),\lim_{\varepsilon\downarrow 0}\bar{\theta}^*(p(\varepsilon))\Big)$$

$$\ge(\rho^*(\theta_T)-\theta)q^f(\rho^*(\theta_T),\theta_T)$$

$$>(\rho^*(\theta_T)-\theta)q^f(\rho^*(\theta_T),\bar{\theta}(T)),$$

[23] Straightforward algebra establishes $\bar{p}_{i-1}>\theta_i$. Thus, $p>\theta_i$. Moreover, without loss of generality, we have focused on the case in which $p\ge\theta_{i+1}$. If $p<\theta_{i-1}$, then $h(\theta_j,p)=\infty$ for all $j\ge i+1$ and the result that we are trying to establish in this part of the proof, $h(\theta_i,p)\le h(\theta_j,p)$ for $j\ge i+1$, follows trivially.

where both inequalities follow because $q^f(p,\theta)$ is decreasing in θ. But (A9) implies that a cartel in the pool can increase profits by charging a slightly lower price, a contradiction of the conditions for a sequential equilibrium. □

PROOF OF PROPOSITION 1:

(a) Suppose, to the contrary, that $p^*(\theta_m) \neq p^f(\theta_m)$. Then,

$$q^*\big(p^f(\theta_m)\big) = q^f\big(p^f(\theta_m), \bar{\theta}^*\big(p^f(\theta_m)\big)\big)$$
$$\geq q^f\big(p^f(\theta_m), \theta_m\big),$$

because $\bar{\theta}^*(p^f(\theta_m)) \leq \theta_m$ and $q^f(p,\theta)$ is decreasing in θ. Thus

$$\text{(A10)} \quad \big(p^f(\theta_m) - \theta_m\big) q^*\big(p^f(\theta_m)\big)$$
$$\geq \big(p^f(\theta_m) - \theta_m\big) q^f\big(p^f(\theta_m), \theta_m\big) \equiv \pi^f(\theta_m),$$

where $\pi^f(\theta_m)$ denotes full-information monopoly profits. But by Lemma 3 the equilibrium is fully separating. Thus $\bar{\theta}^*(p^*(\theta_m)) = \theta_m$, which implies

$$\text{(A11)} \quad \pi^*(\theta_m) = \big(p^*(\theta_m) - \theta_m\big) q^f\big(p^*(\theta_m), \theta_m\big)$$
$$< \pi^f(\theta_m).$$

The inequality in (A11) follows because $p^f(\theta_m)$ is the unique maximizer of $(p - \theta_m)q^f(p, \theta_m)$. But (A10) and (A11) imply that the cartel can increase profits by charging $p^f(\theta_m)$, contradicting the definition of a sequential equilibrium. □

(b) Let us refer to the constrained optimization problem in the statement of the proposition as Problem-(b). Let $\tilde{p}(\theta_i)$ be a solution to Problem-(b). One can show that a solution to Problem-(b) has the following features[24]

$$\text{(A12)} \quad \tilde{p}(\theta_i) \leq p^f(\theta_i)$$

$$\text{(A13)} \quad \text{For } p < \tilde{p}(\theta_i), \ \frac{d}{dp}\big[(p - \theta_i) q^f(p, \theta_i)\big] > 0.$$

$$\text{(A14)} \quad \text{For } p < \tilde{p}(\theta_i), \ (p - \theta_{i+1}) q^f(p, \theta_i)(1 - \beta t) < \pi^*(\theta_{i+1}).$$

$$\text{(A15)} \quad \tilde{p}(\theta_i) \text{ is unique.}$$

If we can show that a sequential equilibrium price $p^*(\theta_i)$ solves Problem-(b), property (A15) implies that the sequential equilibrium must be unique.

Suppose, to the contrary, that the sequential equilibrium price $p^*(\theta_i)$ is *not* the solution to Problem-(b); that is, $\tilde{p}(\theta_i) \neq p^*(\theta_i)$. There are two possibilities: either $\tilde{p}(\theta_i) < p^*(\theta_i)$ *or* $\tilde{p}(\theta_i) > p^*(\theta_i)$.

If $\tilde{p}(\theta_i) < p^*(\theta_i)$, then given the structure of beliefs in Lemma 2, $\bar{\theta}^*(\tilde{p}(\theta_i)) \leq \theta_i$. This implies

$$q^*\big(\tilde{p}(\theta_i)\big) \geq q^f\big(\tilde{p}(\theta_i), \theta_i\big),$$

which, in turn, implies

$$\text{(A16)} \quad \big(\tilde{p}(\theta_i) - \theta_i\big) q^*\big(\tilde{p}(\theta_i)\big)(1 - \beta t)$$
$$\geq \big(\tilde{p}(\theta_i) - \theta_i\big) q^f\big(\tilde{p}(\theta_i), \theta_i\big)(1 - \beta t)$$
$$> \big(p^*(\theta_i) - \theta_i\big) q^f\big(p^*(\theta_i), \theta_i\big)(1 - \beta t)$$
$$= \big(p^*(\theta_i) - \theta_i\big) q^*\big(p^*(\theta_i)\big)(1 - \beta t).$$

The second inequality follows because the equilibrium price $p^*(\theta_i)$ must satisfy (6) (the equilibrium must be incentive compatible) and is thus a *feasible*, though not optimal (given our contrapositive assumption), solution to Problem-(b). But the condition (A16) contradicts the definition of a sequential equilibrium, so we cannot have $\tilde{p}(\theta_i) < p^*(\theta_i)$. If $\tilde{p}(\theta_i) > p^*(\theta_i)$, then (A13) and (A14) are relevant. In particular, (A14) can be shown to imply

$$p^*(\theta_i) < \bar{p}_i,$$

where $\bar{p}_i$ is defined as in Lemma 2. For a price p above $p^*(\theta_i)$ but less than $\bar{p}_i$, $\gamma_i^*(p) = 1$ so $q^*(p) = q^f(p, \theta_i)$ in this neighborhood. But (A13) would then imply that

$$(p - \theta_i) q^*(p)(1 - \beta t)$$
$$> \big(p^*(\theta_i) - \theta_i\big) q^*\big(p^*(\theta_i)\big)(1 - \beta t),$$

for $p \in (p^*(\theta_i), \bar{p}_i)$, contradicting the definition of a sequential equilibrium. □

(c) The specification of equilibrium consumer demand follows directly from the specification of consumer beliefs in Lemma 2. □

PROOF OF PROPOSITION 2: See part (b) of the previous proof. □

PROOF OF PROPOSITION 3:

(a) Note that $\beta t \leq \phi^*$ implies $\beta t \leq \phi_i$ for all $i \in \{1, \ldots, m-1\}$. Straightforward induction arguments establish that constraint (6) will not bind at the optimal solution to Problem-(b), for all $i \in \{1, \ldots, m-1\}$. □

(b) If $\beta t > \phi^*$, there exists some $i \in \{1, \ldots, m-1\}$ such that $\beta t > \phi_i$. Let k be the largest such i. For $j > k$ we thus have welfare neutrality and $\pi^*(\theta_j) = \pi^M(\theta_j)$ for $j > k$. But the Problem-(b) constraint will bind because $\beta t > \phi_k$ implies

$$\pi^M(\theta_k) > \pi^M(\theta_{k+1}) + (\theta_{k+1} - \theta_k) q^M(\theta_k)(1 - \beta t).$$

[24] Proofs of these properties are available on request.

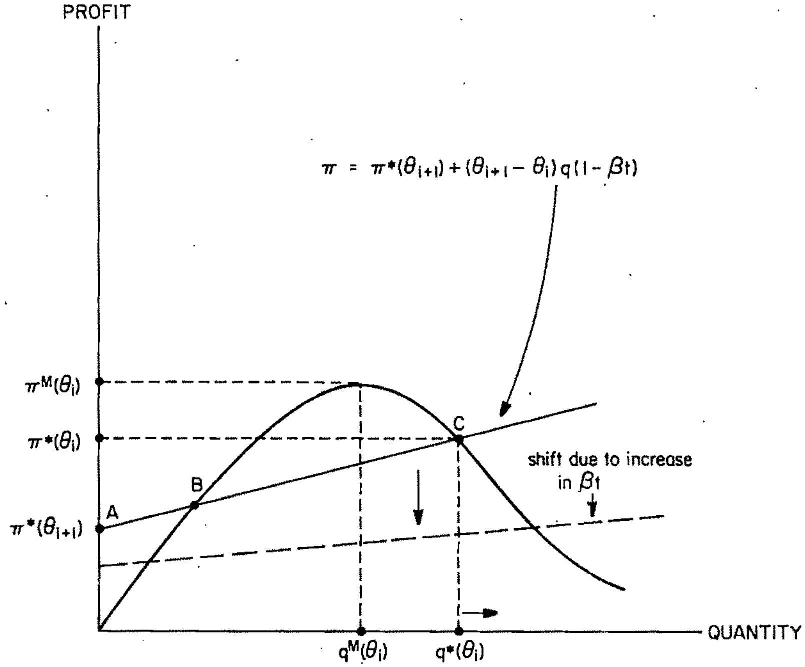

PROFIT
π = π*(θi+1) + (θi+1 − θi)q(1 − βt)
πM(θi)
π*(θi)
π*(θi+1)
A
B
C
shift due to increase in βt
QUANTITY
qM(θi)
q*(θi)

Figure A1. Comparative Statics on Cartel's Constrained Optimization Problem

When constraint (6) binds, welfare neutrality no longer obtains and $p^*(\theta_k) < p^f(\theta_k)$. □

Linear Demand / Uniform Distribution Example. For linear demand and a uniform distribution of cost types

$$\phi_i = \frac{\xi}{2}(a-\theta_i)^{-1},$$

so $\phi_{i+1} > \phi_i$. Find the lowest i such that $\phi_i \geq \beta t$. Label that by the index j. For $i \in \{j,\dots,m-1\}$ condition (7) holds. Arguing by induction, one can prove that the equilibrium is welfare neutral for $i \in \{j,\dots,m\}$. For $i = j-1$, condition (7) does not hold and the equilibrium will not be welfare neutral. Thus,

$$\pi^*(\theta_{j-1}) < \pi^M(\theta_{j-1}).$$

Now we have $\phi_{j-2} < \phi_{j-1} < \beta t$. It therefore follows that

$$\text{(A17)} \qquad \pi^M(\theta_{j-2})$$
$$> \pi^M(\theta_{j-1})$$
$$+ (\theta_{j-1}-\theta_{j-2})q^M(\theta_{j-2})(1-\beta t).$$
$$> \pi^*(\theta_{j-1})$$
$$+ (\theta_{j-1}-\theta_{j-2})q^M(\theta_{j-2})(1-\beta t).$$

Examination of Problem-(b) reveals that when (A17) holds, the equilibrium must result in a welfare improvement; that is, $p^*(\theta_{j-2}) < p^f(\theta_{j-2})$. Repeating this logic for $j-3,\dots,1$ implies that the equilibrium results in a welfare improvement for all $i \in \{j-1,\dots,1\}$. □

PROOF OF PROPOSITION 4: Using the expressions for $q^f(p,\theta_i)$ and $p^f(q,\theta_i)$ from Section II, we can restate Problem-(b) as

$$\pi^*(\theta_i) = \max_q \left(P(q)-\theta_i\right)q$$

$$\text{subject to: } \left(P(q)-\theta_i\right)q \leq \pi^*(\theta_{i+1})$$
$$+(\theta_{i+1}-\theta_i)(1-\beta t)q.$$

As a first step, we show that

$$\text{(A18)} \qquad \frac{d\pi^*(\theta_i)}{d(\beta t)} \leq 0 \quad \text{for all } i.$$

The proof is by induction. Condition (A18) holds for $i = m$ (since $\pi^*(\theta_m) = \pi^M(\theta_m)$, so $d\pi^*(\theta_m)/d(\beta t) = 0$). Suppose (A18) holds for $i = j+1$. We will now prove that it also holds for $i = j$, which establishes the result. By the envelope theorem

$$\frac{d\pi^*(\theta_j)}{d(\beta t)} = \lambda_j\left\{\frac{d\pi^*(\theta_{j+1})}{d(\beta t)} - (\theta_{i+1}-\theta_i)q^*(\theta_j)\right\},$$

where $\lambda_j \geq 0$ is the Lagrange multiplier for the constraint in Problem-(b) and $q^*(\theta_j)$ is the solution. Clearly, $d\pi^*(\theta_j)/d(\beta t) \leq 0$, so, by induction, the result holds for all i.

Next, suppose that the equilibrium is not welfare neutral at θ_i. Then the solution to Problem-(b) is as shown in Figure A1. Given (A18), an increase in βt shifts the line segment ABC downward and flattens it out. It follows that $q^*(\theta_i)$ increases and $\rho^*(\theta_i)$ decreases. □

REFERENCES

Baker, Jonathan B., "Private Information and the Deterrent Effect of Antitrust Damage Remedies," *Journal of Law, Economics and Organization*, Fall 1988, *4*, 385–408.

Banks, Jeffrey S., "Regulatory Auditing Without Commitment," Working Paper, University of Rochester, 1988.

______ **and Sobel, Joel,** "Equilibrium Selection in Signalling Games," *Econometrica*, May 1987, *55*, 647–61.

Besanko, David and Spulber, Daniel F., (1989a) "Antitrust Enforcement Under Asymmetric Information," *Economic Journal*, June 1989, *99*, 408–25.

______ **and** ______, (1989b) "Delegated Law Enforcement and Noncooperative Behavior," *Journal of Law, Economics and Organization*, Spring 1989, *5*, 25–52.

Breit, William and Elzinga, Kenneth G., "Antitrust Enforcement and Economic Efficiency: The Uneasy Case for Treble Damages," *Journal of Law and Economics*, October 1974, *17*, 329–56.

______ **and** ______, "Private Antitrust Enforcement: The New Learning," *Journal of Law and Economics*, May 1985, *28*, 405–43.

Coase, Ronald H., "The Problem of Social Cost," *Journal of Law and Economics*, October 1960, *3*, 1–44.

Easterbrook, Frank H., "Detrebling Antitrust Damages," *Journal of Law and Economics*, May 1985, *28*, 445–67.

Elzinga, Kenneth G. and Breit, William, *The Antitrust Penalties: A Study in Law and Economics*, New Haven, CT.: Yale University Press, 1976.

Kreps, David M. and Wilson, Robert, "Sequential Equilibria," *Econometrica*, July 1982, *50*, 863–94.

Mailath, George J., "Incentive Compatibility in Signalling Games with a Continuum of Types," *Econometrica*, November 1987, *55*, 1349–66.

Polinsky, A. Mitchell, "Detrebling Versus Decoupling Antitrust Damages: Lessons from the Theory of Enforcement," *Georgetown Law Journal*, April 1986, *74*, 1231–36.

Posner, Richard A., *Antitrust Law: An Economic Perspective*, Chicago: University of Chicago Press, 1976.

Salant, Steven W., "Treble Damage Awards in Private Lawsuits for Price Fixing," *Journal of Political Economy*, December 1987, *95*, 1326–36.

Salop, Steven C. and White, Lawrence J., "Private Antitrust Litigation: An Introduction and Framework," in L. J. White, ed., *Private Antitrust Litigation: New Evidence, New Learning*, Cambridge, MA: MIT Press, 1988, ch. 1.

Spiller, Pablo T., "Treble Damages and Optimal Suing Time," *Research in Law and Economics*, April 1986, *9*, 45–56.

Spulber, Daniel F., "Contract Damages and Competition," Working Paper, University of Southern California, 1987.

______, *Regulation and Markets*, Cambridge, MA.: MIT Press, 1989.

Schwartz, William F., "An Overview of the Economics of Antitrust Enforcement," *Georgetown Law Journal*, June 1980, *68*, 1075–1102

Ex Post Liability for Harm vs. *Ex Ante* Safety Regulation: Substitutes or Complements?

By CHARLES D. KOLSTAD, THOMAS S. ULEN, AND GARY V. JOHNSON*

This paper concerns the regulation of hazardous economic activities. Economists have generally viewed ex ante regulations (safety standards, Pigouvian fees) that regulate an activity before an accident occurs as substitutes for ex post policies (exposure to tort liability) for correcting externalities. This paper shows that where there is uncertainty, there are inefficiencies associated with the exclusive use of negligence liability and that ex ante regulation can correct the inefficiencies. In such a case it is efficient to set the safety standard below the level of precaution that would be called for if the standard were used alone. (JEL *619*)

One of the main issues that dominates the economic literature on optimal regulation is the choice of the most efficient policy for correcting an externality. From its beginnings the literature has focused on alternative forms of what may be called *ex ante* policies (for example, safety standards, Pigouvian taxes, and transferable discharge permits) that affect an activity before the externality is generated. But in the past decade researchers have analyzed the ability of what may be called *ex post* policies (for example, exposure to tort liability) to control externalities.[1] These latter policies regulate the externality only after it has been generated and harm has occurred. The threat of suit causes the potential injurer to internalize the expected social damages and thus to take optimal precaution.

Economists have generally viewed *ex ante* and *ex post* policies as substitutes for correcting externalities. The usual policy recommendation has been to choose the less costly regulatory policy to administer. For instance, in the case of chopping down a tree in one's yard, it is less costly to use threat of suit to force appropriate caution than to construct a myriad of permits and regulations covering tree-falling. An example at the other extreme is air pollution, where it is less costly to promulgate well-thought-out regulations than to let each injured party take injurers to court. Rarely is the joint use of *ex ante* and *ex post* policies recommended for a given externality.[2]

This theoretical conclusion, however, stands in stark contrast to actual policy. One of the most noticeable features of current policy dealing with externality-generating activities in a wide number of areas is that *ex ante* and *ex post* policies are very frequently used jointly. Consider the following examples. The potential inefficiencies of incompatible neighboring property uses—for example, a hospital located next to a noisy, dusty cement-manufacturing plant—are minimized by zoning ordinances (a form of *ex ante* regulation) and by simultaneously exposing the externality generator to nui-

*Department of Economics and Institute for Environmental Studies, University of Illinois, Urbana, IL 61801; Department of Economics and College of Law, University of Illinois, Champaign, IL 61820; Food Marketing Policy Center, Department of Agricultural Economics, University of Connecticut, Storrs, CT 06268. We have benefited from comments by two anonymous referees, Robert Cooter, Richard Craswell, Robert Gritz, Judy Lachman, Steven Shavell, V. Kerry Smith, and participants in workshops at MIT, Colby, and the Universities of British Columbia, Chicago, Illinois, and Washington.

[1] John Brown (1973) and Peter Diamond (1974) were among the first to mathematically articulate Guido Calabresi's (1970) theories of liability as a means of controlling externalities.

[2] An exception is the recent work of Steven Shavell (1984a,b), which is discussed in more detail later in this paper.

sance liability (a form of *ex post* regulation).[3] Similarly, society attempts to minimize the harms that new pharmaceuticals may inflict on users by requiring the manufacturers of drugs to engage in specific tests before the drugs are licensed by the federal Food and Drug Administration for prescription and sale (a form of *ex ante* regulation) and also by thereafter exposing the drug manufacturers to strict products liability (*ex post* regulation). In the field of environmental externalities, the potential harms of toxic wastes are regulated at the federal level by the Resource Conservation and Recovery Act (1982), which imposes *ex ante* siting and technological regulations on the generation and disposal of hazardous wastes, and the Comprehensive Environmental Resource, Compensation, and Liability Act of 1979, which establishes *ex post* liability rules for the recovery of compensatory and punitive damages for harms imposed by hazardous wastes.

This phenomenon of complementary use of *ex ante* and *ex post* regulatory policies is so widespread that the dearth of persuasive theoretical arguments for this joint use is glaring. Various authors have identified inefficiencies associated with one or the other regulatory policy. In the case of *ex ante* regulation, the typical criticism is that the central regulator has imperfect information on accident costs and damages (Martin Weitzman, 1974; William Baumol and Wallace Oates, 1971; Susan Rose-Ackerman, 1973; Shavell, 1984b), which may lead to inefficient undercontrol of some wrongdoers and overcontrol of others. The typical criticisms of tort liability have been that suit may not always be brought against injurers, that bankruptcy provides an incentive for underprotection, and that uncertainty regarding the legal standard leads to over- or underprotection, depending on the circumstances.[4] Shavell (1984b, 1987) appears to be alone in suggesting that *ex ante* and *ex post* regulation can complement each other in that their joint use can correct the inefficiencies of using either alone to correct an externality.

This paper first identifies a set of inefficiencies associated with *ex post* liability. These inefficiencies are due to a potential injurer's being uncertain about whether a court will hold him liable in the event of an accident and suit. Our discussion formalizes and extends the results and conjectures of Craswell and Calfee (1985) and Calfee and Craswell (1984). In contrast to Shavell (1984b, 1987), we do not base our analysis upon the inefficiencies due to the potential bankruptcy of injurers and the uncertainty of suit by victims. Nor do we assume risk aversion. Having identified inefficiencies associated with tort liability, we then demonstrate how *ex ante* regulation, if used jointly with tort liability, can correct some of those inefficiencies.

One of our strongest conclusions, and a startling one, is that when *ex ante* and *ex post* policies should be used jointly, efficiency generally requires that the *ex ante* regulatory standard be set on a level that, if regulation were used alone, would provide a socially *suboptimal* level of safety or precaution. Put somewhat differently to emphasize this unconventional conclusion, when tort liability rules are in place, it is *inefficient* to set *ex ante* regulatory standards at the socially optimal level (where marginal costs of precaution equal the expected marginal benefits). The only instances when the *ex ante* regulatory standard should be set at the social optimum are when there is no *ex post* liability or, equivalently, when there is a zero probability of a judgment against a rational injurer under *ex post* liability.

I. A Model of Negligence and Safety Regulation

Consider the case in which a risk-neutral firm (or any other economic agent) engages in a risky activity, from which accidents can occur. The firm can reduce the dangers associated with this activity by taking precaution. Precaution reduces expected accident costs but is costly to the firm.

[3] The classic comparison of the efficiency aspects of these alternate methods of minimizing this type of externality is given by Robert Ellikson (1973).

[4] See Brown, 1973; Robert Cooter et al., 1979; Richard Craswell and John Calfee, 1986; Shavell, 1984b; 1987; Donald Wittman, 1977.

Let x be the level of the firm's precaution in preventing an accident or reducing its severity. For simplicity, we will not consider the decisions of the potential victim by assuming she always takes the socially optimal level of precaution.[5] The injuring firm's costs of taking precaution are given by the function $C(x)$, which is upward sloping $[C'(x) > 0]$ and convex over the relevant region. An accident will occur with probability $p(x, \varepsilon)$ and will be of size (cost) $D(x, \varepsilon)$ where ε is a random variable representing the view-of-the-court and is distributed with density function q_ε. Assume the expected value of ε is zero. By this assumption we mean to suggest that, on average, courts evaluate the precautionary behavior of the injurer and the size of the accident $(D(\cdot))$ accurately. Define $A(x)$ as the expectation of $p(x, \varepsilon)D(x, \varepsilon)$ over ε. Thus, $A(x)$ embodies both the accident size (D) and the probability of the accident occurring (p). The view-of-the-court is only revealed after a court has heard evidence about x and the extent of damage after an accident has occurred. Assume $A(x)$ is convex and downward sloping over the relevant region $(A'(x) < 0)$. Assume that $(C(x)+A(x))$ is strictly convex.

To avoid confusion, it is useful to preview the three fundamentally different levels of precaution we will consider. We first define the socially optimal amount of precaution, x^*, where the expected social costs of accidents are minimized. We then define the legal standard of care, $\bar{x}(\varepsilon)$, which is the court's interpretation of the social optimum, a function of ε since it is only revealed after an accident and litigation occurs. The third type of precaution is the firm's precaution level, $\hat{x}$, chosen to minimize expected private costs to the firm. Our goal will be to compare $\hat{x}$ and x^*.

The socially optimal amount of precaution for the potential injurer can be obtained by minimizing expected social costs; that is,

$$(1) \quad \min_x E[C(x) + p(x, \varepsilon)D(x, \varepsilon)] = \min_x [C(x) + A(x)].$$

At the unique level of x that minimizes equation (1), x^*, the marginal cost of precaution equals the negative of the marginal expected cost of the accident; that is,

$$(2) \qquad C'(x^*) = -A'(x^*),$$

assuming the solution of equation (1) is greater than zero.

The legal standard, as opposed to the social optimum, is an *ex post* parameter, revealed by the courts after an accident has occurred. Thus the legal standard is parameterized by the view-of-the-court: $\bar{x}(\varepsilon)$. In fact $\bar{x}(\varepsilon)$ is defined as the solution of

$$(3) \quad \min_x [C(x) + p(x, \varepsilon)D(x, \varepsilon)]$$

for which the first-order condition is

$$(4) \quad C'(x) + \frac{d[p(x, \varepsilon)D(x, \varepsilon)]}{dx} = 0,$$

assuming an interior minimum. Equation (4) implicitly defines $\bar{x}(\varepsilon)$. Since ε is a random variable, $\bar{x}(\varepsilon)$ induces a distribution on $\bar{x}$, which we term $q_{\bar{x}}$ or more simply, q. We assume q is continuous.

It is at this point that the notion of liability enters. Under a negligence rule, the injurer is found liable for all damages if, and only if, his level of precaution was less than the legal standard of precaution. Mathematically, the injurer's total expected costs are given by

$$(5) \quad TC(x) = E[C(x) + L(x, \varepsilon) \times p(x, \varepsilon)D(x, \varepsilon)],$$

where $L(x, \varepsilon)$, the liability rule, is defined

[5] Many authors (for example, Diamond, 1974; Brown, 1973; Cooter et al., 1979) explicitly consider the level of precaution taken by the potential injured party. While this is realistic and leads to richer conclusions in many analyses, it is tangential to the purposes of this paper.

by

$$\text{(6) Negligence: } L(x,\varepsilon) = \begin{cases} 1 & \text{if} \quad x < \bar{x}(\varepsilon) \\ 0 & \text{otherwise} \end{cases}.$$

Let $\tilde{x}$ be the level of precaution that minimizes (5). Ideally $\tilde{x}$ should equal x^*, in which case the liability rule is *ex ante* efficient.

There is some potential controversy about what we mean by the legal standard's being parameterized by the court's view. We mean simply that at the moment at which the potential injurer makes her decision about the appropriate level of precaution, she faces some unavoidable uncertainty about how a fact finder will subsequently evaluate her decisions. She does not know, for instance, whether the court will evaluate her actions according to an objective standard (for example, what precaution would a reasonable person have taken in these circumstances?) or a subjective standard (for example, will the court attempt to place itself in the decision environment that the injurer was in when she decided how much precaution to take?) (Calabresi and Alvin Klevorick, 1985). In addition, even if there is no uncertainty about how the court will look at the injurer's precaution, there may be uncertainty about the court's ability to evaluate the evidence of the injurer's precaution. The injurer and the court may sample from the same population, but their sample means need not be identical.[6]

A basic result of Brown (1973) is that when the legal standard is defined as in (4) and the firm knows that standard with certainty, then the negligence rule is efficient. This conclusion in the case of negligence is qualified by Calfree and Craswell (1984). Their argument hinges on *ex ante* uncertainty on the part of the firm regarding the legal standard, $\bar{x}(\varepsilon)$. Unfortunately, for the most part Calfee and Craswell are unable to prove their conjectures and must rely on cogent argument and numerical examples.[7]

Shavell (1984b, 1987) provides the only thorough treatment of correcting the inefficiencies of tort liability by supplementing it with *ex ante* safety regulation. But instead of relying on uncertainty, Shavell argues that negligence is inefficient because $L(x) < 1$. And this, he suggests, is due to (a) a positive probability that suit will never be brought against an injurer, and (b) an injurer not needing to plan for accidents whose cost (D) exceeds the injurer's assets.

For symmetry, Shavell also suggests that *ex ante* regulation by itself is inefficient because D is not known with certainty to the regulator (but is known by the injurer) and pure *ex ante* regulation requires one level of care for all injurers. This means that firms that cause small accidents are overregulated and firms that cause large accidents are underregulated. A mixed regulatory system

[6] These forms of uncertainty make our fundamental point that an injurer, *ex ante*, will be unsure about a court's interpretation or view of the evidence. Consider further the uncertainties that arise in an injurer's mind about whether she will be held to some form of the negligence standard (and what form?) or to a strict liability standard (under which the victim's precautionary behavior is not evaluated), about what jurisdiction and under which judge litigation may occur, and about whether she might be held liable for punitive damages or just for compensatory damages. Finally, as Judge Posner (1988) has noted, there is an inherent uncertainty about cases litigated to judgment because, if the matter were clear, the litigants would have settled out of court. Diamond (1974) views this uncertainty from a somewhat different perspective. He assumes the firm knows the legal standard of care with certainty but the firm is uncertain about how its precautionary measures translate into safety levels and it is these safety levels that are measured by the court. However, the effect is the same: for a given level of precaution the firm is uncertain as to whether it is above or below the legal standard. Cooter and Tom Ulen (1986) examine evidentiary uncertainty, or uncertainty in exactly how a court will interpret evidence in deciding whether the firm's level of precaution was above or below the "legal standard."

[7] Craswell and Calfee (1986) prove that for a legal standard symmetrically distributed about x^*, small levels of uncertainty lead to oversupply of precaution, provided density is concentrated at x^* for low levels of uncertainty. At the optimal level of care, the marginal costs of precaution just offset the marginal accident costs from precaution. But the injurer also sees a marginal savings in liability due to the unpredictability of the legal standard. Thus, the potential injurer may take precaution $\tilde{x} \neq x^*$.

results in firms that cause little damage being regulated by the *ex ante* regulation and firms that cause great damage being regulated by the threat of liability. Given the inefficiencies built into *ex ante* safety regulation and *ex post* liability for harm, it is easy to show that a hybrid does no worse and frequently does better than either approach alone.[8] Key to the Shavell analysis is that $L(x)$ is strictly less than 1. If bankruptcy is not a possibility and suit is always brought, then $L(x)=1$ and there are no inefficiencies associated with liability.

We take a different approach in this paper. Similarly to Craswell and Calfee (1986), we suggest that it is uncertainty over the legal standard that leads to inefficiencies with negligence. In fact, in the next section we prove all of their conjectures as well as others regarding the efficiency of negligence liability. In the subsequent section of the paper, we show how *ex ante* regulation can be used to correct some of these inefficiencies.

II. The Inefficiency of Negligence

Our basic model of negligence was developed in the previous section. The legal standard of precaution, $\bar{x}(\varepsilon)$, is defined implicitly by (4). Should an accident occur, litigation will reveal the true view-of-the-court, ε, and its realization, $\bar{x}$, on q. If a court finds that the firm's level of precaution was less than $\bar{x}$, then the firm will be liable for all accident costs; if greater than $\bar{x}$, no liability will apply.

The firm does not know the view-of-the-court, ε, when it chooses $\tilde{x}$. The firm must choose an $\tilde{x}$ based on an uncertain legal standard, $\bar{x}(\varepsilon)$. As defined above, we let $q(x)$ be the injurer's subjective probability distribution around the legal standard, the level of precaution that the firm must provide to avoid being held liable for accident costs. We assume that $q(x)$ is a continuous probability density with support $(-\infty,\infty)$.[9] The probability that the injurer's level of precaution x will end up being below the legal standard of care applied in the case of an accident is thus given by

$$(7)\qquad R(x)=\int_{x}^{\infty} q(x)\,dx.$$

That is, $R(x)$ is the probability when all is said and done, after the court has passed judgment, that the injurer will pay damages $E[p(x,\varepsilon)D(x,\varepsilon)]=A(x)$. We have already assumed that $C(x)$ and $A(x)$ are convex. We now make the slightly stronger assumption that $[C(x)+A(x)R(x)]$ is strictly convex.

The essence of our model is presented in Figure 1. The expected legal standard, which is defined to be the socially optimal level of precaution, is where the marginal precaution costs just equal the negative of the marginal expected accident costs, as indicated in equation (2). With uncertainty, the injurer does not know $\bar{x}$ precisely. The injurer's uncertainty about the legal standard to which it will be held accountable is embodied in $q(x)$. If the injurer takes $\tilde{x}$ amount of precaution, then the probability that it will be held liable is the area under the density function from $\tilde{x}$ to ∞, $R(\tilde{x})$, the cross-hatched area in the figure. This is the probability that should an accident occur, the firm will be found to be taking an inadequate amount of precaution and, thus, be liable for the victim's harm.

When there is this type of uncertainty, the injurer's objective function (in the presence of potential liability) is defined by (5) except that $L(x)$ is replaced by $R(x)$:

$$(8)\quad TC(x)=E[C(x)+R(x)p(x,\varepsilon)D(x,\varepsilon)]=C(x)+A(x)R(x),$$

[8] This result is strikingly similar to that of Marc Roberts and Michael Spence (1976). They argue, in an entirely different context, for a hybrid system of price and quantity controls to optimally control an externality. The analogy to our problem is that quantity controls are akin to *ex ante* regulations and price controls are similar to tort liability (in that liability induces the firm to equate marginal damage and marginal precaution costs).

[9] Alternatively, one could argue that the support should be $[0,\infty]$.

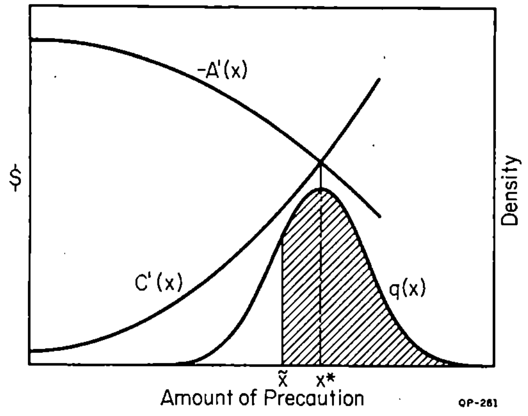


FIGURE 1. THE SOCIAL PROBLEM WITH EVIDENTIARY UNCERTAINTY FOR THE INJURER

which the firm attempts to minimize. $TC(x)$ is strictly convex, by assumption, and thus has a unique minimum. Let $\tilde{x}$ be the level of precaution that minimizes (8). The first-order condition for the minimization is

$$(9)\quad TC'(\tilde{x}) = C'(\tilde{x}) + A'(\tilde{x})R(\tilde{x}) - A(\tilde{x})q(\tilde{x}) = 0,$$

provided $\tilde{x}$ is greater than zero.

Equation (9) is basic to much of our analysis and thus deserves some interpretation. The first term on the right-hand side of equation (9) is the marginal cost of providing the $\tilde{x}$th unit of precaution, and we assume that this marginal cost is rising. The second and third terms, which together we might call "marginal (expected) liability costs," sum to the expected marginal liability costs of the $\tilde{x}$th unit of precaution and consist of two effects. The first of these terms $[A'(\tilde{x})R(\tilde{x})]$ is the marginal (reduction in) accident cost times the probability of being held liable for the accident if the firm has taken precaution equal to $\tilde{x}$. This term, which might be called the "injury effect," is negative because $A'(\tilde{x})$ is negative and $R(\tilde{x})$ is always positive. The injury effect represents a savings to the injurer from the application of greater precaution because expected accident costs are reduced. But there is also a savings from providing slightly higher precaution in that the probability of being held liable is reduced. The monetary savings is the product of the change in the probability of liability and total expected accident costs. This savings is captured in the term $[-A(\tilde{x})q(\tilde{x})]$. This term, which might be called the "liability effect," is negative because both $A(\tilde{x})$ and $q(\tilde{x})$ are positive.[10] Thus, the marginal (expected) liability costs can be decomposed into an injury effect and a liability effect, both of which have the same sign and both of which indicate that marginal (expected) liability costs decline in precaution.

The question that arises is whether the level of precaution, $\tilde{x}$, chosen by the firm to minimize its expected costs is greater than, less than, or equal to the socially optimal level of precaution, x^*. An evaluation of the relationship between $\tilde{x}$ and x^* can be made by evaluating the sign of $TC'(x^*)$ in equation (9). Since by assumption $\tilde{x}$ minimizes $TC(x)$ and $TC(x)$ is strictly convex, $TC'(x) < 0$ for $x < \tilde{x}$ and $TC'(x) > 0$ for $x > \tilde{x}$. Thus, if $TC'(x^*) < 0$, then $x^* < \tilde{x}$; and if $TC'(x^*) > 0$, then $x^* > \tilde{x}$. Evaluating $TC'(x)$ (in equation (9) at x^*, using equation (2), gives

$$(10)\quad TC'(x^*) = C'(x^*)[1 - R(x^*)] - A(x^*)q(x^*).$$

Since $C'(x) \geq 0$ by assumption and $R(x) \leq 1$, then $C'(x^*)[1 - R(x^*)]$ in equation (10) is nonnegative. Also, since by definition $A(x)$ and $q(x)$ are greater than or equal to zero, the term $-A(x^*)q(x^*)$ is nonpositive. Therefore, the sign of equation (10) is indeterminate and the relationship between $\tilde{x}$ and x^* cannot be discovered without knowing the magnitude of the various terms. Any further evaluation of equation (10) will require further assumptions regarding the nature of the distribution $q(x)$ and the size of the marginal cost of precaution. First let us turn to assumptions about the nature of the distribution $q(x)$.

[10] The negative sign in front of the term $[A(\tilde{x})q(\tilde{x})]$ is due to the fact that $R'(\tilde{x}) = -q(\tilde{x})$.

A. *The Effect of the Level of Uncertainty About the Legal Standard*

We are concerned here with the extent of the uncertainty regarding $\tilde{x}$; for example, the size of the variance of $\tilde{x}$, which is distributed as $q(x)$. Let us consider two cases, one where there is a great deal of uncertainty with regard to the legal standard, and one where there is little uncertainty with regard to the standard. An example of the first case is the great uncertainty regarding the appropriate standard of care when dealing with a new technology, for example, genetic engineering. The level of scientific knowledge regarding the potential for accidents and the extent of the damages may be low, and it may therefore be difficult to determine in the first instances of accidents what the socially optimal level of precaution in testing, production, warnings, and disposal, is for genetically engineered output. An example of the second case, where there is very little uncertainty about the appropriate legal standard, might be the case for a well-recognized harm where the costs and benefits of accident precaution are well known and legal precedent is well established, for example, an automobile accident.

We are concerned with the effect of the extent of uncertainty in $\tilde{x}$, embodied in $q(x)$, on the sign of $TC'(x^*)$ in (10). Before addressing this concern, we must be somewhat more precise about what we mean by more or less uncertainty. The conventional notion is that of second-order stochastic dominance (Steven Lippman and John McCall, 1981). But just because one distribution dominates another in this sense does not assure us that the density function will be any different at x^*, which is important in the present analysis. To facilitate our comparative statics analysis, we will introduce a class of distributions, members of which differ in terms of location and scale parameters (for example, Jack Meyer, 1986); that is, two distributions, with continuous density functions f and g, belong to the same class if there exists an $\alpha > 0$ and β such that

$$(11) \qquad f(x) = \alpha g(\alpha x + \beta).$$

Consider the class of distributions differing only in location and scale (a) that contains $q(x)$, and (b) whose members have identical means, x^*. This defines a family of mean-preserving spreads and implies that $\beta = (1-\alpha)x^*$. In particular, parameterize the set of distributions by

$$(12) \qquad q_\alpha(x) = \alpha q[\alpha x + (1-\alpha)x^*],$$

where (by assumption) the legal standard is distributed as $q(x)$, defined over the nonnegative reals, with expected value x^*. It can readily be seen that $q_1(x) \equiv q(x)$. Furthermore, q_α is a well-behaved density function for all values of $\alpha > 0$, and random variables distributed according to q and q_α have the same mean. As α decreases, the spread of q_α increases; as α increases, the probability mass becomes concentrated at the mean. Thus, in this instance uncertainty in the legal standard is inversely related to the size of α.

As uncertainty becomes larger, that is, as α becomes smaller, $q(x^*)$ becomes smaller, and $TC'(x^*)$ in equation (10) eventually becomes positive. This implies that $\tilde{x}$ is less than x^*. That is, the greater the uncertainty in the legal standard, the more likely it is that a potential injurer will take *less* than the socially optimal amount of precaution. As uncertainty becomes less, that is, as α becomes larger, the probability mass becomes concentrated at x^* and $TC'(x^*)$ becomes negative. This implies that $\tilde{x}$ is greater than x^*. That is, the less the uncertainty in the legal standard, the more likely that the firm will take *more* precaution than the social optimum.

We can now state our first result regarding the effect of uncertainty on the use of negligence as a liability rule.[11]

PROPOSITION 1: *Assume that to an injurer, the legal standard is uncertain (distrib-*

[11] This result for overprotection was proved, for the special case of symmetric distributions, by Craswell and Calfee (1986). Shavell (1987) provides an elegant proof that sufficiently small uncertainty leads to overprotection.

uted as $q(x)$) with an expected value of x^*; assume $q(x^*) > 0$. If uncertainty regarding the legal standard (in the sense of equation (11)) *is sufficiently large (small), then the injurer subject to a negligence rule will underprotect (overprotect).*

PROOF:

To prove that with sufficiently large uncertainty a firm will underprotect, we need to show that there exists a more spread-out version of $q(x)$ such that $TC'(x^*)$ in equation (10) becomes positive. Equation (10) can be rewritten using q_α from equation (12) as

$$(13)\quad TC'(x^*) = C'(x^*)[1 - R_\alpha(x^*)] - A(x^*)q_\alpha(x^*) = C'(x^*)[1 - R_\alpha(x^*)] - A(x^*)\alpha q(x^*).$$

Obviously, there exists an $\alpha > 0$ such that equation (13) is positive (since $C'(x^*) > 0$ and regardless of α, $R_\alpha(x^*) < 1$). Conversely, since $A(x^*) > 0$ (because $A'(x^*) < 0$ and $A(x^*) \geq 0$) and $q(x^*) > 0$, there exists an $\alpha > 0$ such that equation 13 is negative. □

Thus, even though the injurer's expected value of the legal standard is equal to the social optimum, uncertainty is sufficient to result in over- or underprotection.[12] This result flows from the definition of negligence. Under that liability standard, if one slightly undercomplies with the legal standard, one is just as liable as if one greatly undercompiled.[13] In deciding whether to increase or decrease precaution from x^*, the injurer must trade off the marginal cost of precaution against the expected marginal benefits (in the form of the sum of expected marginal accident costs and the change in the likelihood of being found liable). When there is very little uncertainty in the determination of the legal standard, overcompliance results in only slightly higher expenditures for precaution but in a greatly reduced likelihood of being held liable and, therefore, paying for the victim's losses. Similarly, where there is a great deal of uncertainty surrounding the legal standard of care, undercompliance greatly reduces precautionary costs while only slightly increasing expected liability costs.

B. *The Effect of the Marginal Cost of Precaution at the Social Optimum*

We now consider the effect of $C'(x)$ on over- or underprotection, holding $q(x)$ constant. Of course, changes in $C'(x)$ do change x^*. Nevertheless, in equation (10), if marginal costs of protection are sufficiently large, then $TC'(x^*)$ can be driven positive. In that case there can be underprotection, that is, $\tilde{x} < x^*$. By a similar argument it is clear that as $C'(x^*)$ goes to zero, then $TC'(x^*)$ becomes negative, implying overprotection, that is, $\tilde{x} > x^*$.

PROPOSITION 2: *Under the assumptions of Proposition* 1, *if $q(x^*) > 0$, then for a sufficiently small (large) marginal cost of precaution at the social optimum, x^*, the injurer will employ too much (little) precaution to prevent an accident when faced with a negligence rule.*

[12] An alternative explanation of this result can be made with reference to equation (9). The last two terms of equation (9), the marginal liability costs, are the savings from increased precaution. Equation (9) can be rewritten as

$$C'(\tilde{x}) = -A'(\tilde{x})R(\tilde{x}) + A(\tilde{x})q(\tilde{x}).$$

An increase (decrease) in the injurer's uncertainty at a point $\tilde{x}$ decreases (increases) the right-hand side of this equation. This comes about because both $R(\tilde{x})$ and $q(\tilde{x})$ become smaller (larger) as uncertainty increases (decreases). At the equilibrium this implies that a lower (higher) level of precaution is taken by the injurer.

[13] Save, of course, for the possibility that undercompliance that is "gross, malicious, willful and wanton, reckless, or fraudulent" may trigger punitive damages.

The intuition of this proposition is straightforward. If the marginal cost of precaution around the legal standard is very low, then by overcomplying the potential injurer raises his precautionary costs only slightly but greatly reduces his expected liability costs. On the other hand, if the marginal cost of precaution at the legal standard is very large, then undercompliance greatly reduces precautionary costs while increasing expected liability costs relatively slightly.

C. *The Effect of Biased Perceptions of the Legal Standard*

In the previous section we focused on the effect of uncertainty with respect to the legal standard on over-or underprotection. By assumption, the mean of the distribution was the social optimum; what we examined was the effect of changing the variance or spread of the distribution. We now introduce a bias in the firm's perception of the legal standard, $\bar{x}(\varepsilon)$. We are now concerned less with the spread of the distribution than with the extent to which the firm views the distribution of $\bar{x}$ as biased to one side or the other of the social optimum. In particular, we consider the case where the bulk of the probability mass is either to the left or to the right of x^* $(E[\bar{x}(\varepsilon)] \lesseqgtr x^*)$.[14] Note that we are considering the effect of bias in the perception of x^* only on the part of potential injurers. It is difficult to imagine why this should occur. But imagine that firms *believe*, whether rightly or wrongly, that juries consistently make biased assessments of firms' precautionary efforts and of the size of victims' losses (Cooter and Ulen, 1986).

Consider the family of distributions, containing $q(x)$, that differ only in location and scale, as in equation (11), and that share the same variance. This implies that the distributions are shifted with respect to one another:

$$q_\beta(x) = q(x+\beta). \tag{14}$$

Obviously, $q_0 = q(x)$. Thus for $\beta < 0$, the distribution is biased to the right; and for $\beta > 0$, it is biased to the left. Note that if $q(x)$ can be made arbitrarily close to zero for sufficiently small or large x, then by choosing β, $q_\beta(x^*)$ can be made arbitrarily small with R_β arbitrarily close to 0 or 1, depending on whether β is small or large. Clearly if q is a continuous density (as we have assumed), with support $(-\infty, \infty)$, then $q(x)$ must become arbitrarily small for small or large x.

The case where q_β is biased to the right might be the case of a work-related harm in which it is difficult to show causality, for example, an increased incidence of lung disease as a result of a firm's precautionary behavior two decades earlier. Conversely, suppose the firm significantly underestimates the expected legal standard. Then $R(x^*) \approx 1$. This might be the case in emotionally charged accidents where juries may have sympathy for victims.

PROPOSITION 3: *If the distribution $q(x)$ is sufficiently biased to the left (right) of x^* in the above sense, then the injurer will underprotect (overprotect) against an accident when faced with a negligence rule.*

PROOF:

To prove that if the distribution is sufficiently biased one gets over-or under protection, we parameterize $q(x)$ as in equation (14). Assume x distributed as $q(x)$ has mean x^*. Equation (10) then becomes

$$\begin{aligned} TC'(x^*) &= C'(x^*)\left[1 - R_\beta(x^*)\right] \\ &\quad - A(x^*)q_\beta(x^*) \\ &= C'(x^*)\left[1 - R_\beta(x^*)\right] \\ &\quad - A(^*)q(x^*+\beta). \end{aligned} \tag{15}$$

Since $R(x) \to 0$ as $x \to \infty$, clearly $R_\beta(x^*) \equiv R(x^*+\beta) \to 0$ ad $\beta \to \infty$. By continuity of $q(x)$, this implies that $q(x^*+\beta) \to 0$ as β

[14] This is a generalization and extension of the case considered by Craswell and Calfee (1986) of the whole distribution shifting to the right or left, although our results support their conjectures.

$\rightarrow \infty$. Thus there exists a sufficiently large β for which $TC'(x^*)$ in equation (15) becomes positive, implying that $\tilde{x} < x^*$. Therefore, for uncertainty sufficiently biased to the left, injurers will underprotect under a negligence rule. Trivially, as $\beta \rightarrow -\infty$, then $TC'(x^*)$ becomes negative: if uncertainty is sufficiently biased to the right, injurers will overprotect. □

This proposition has a straightforward interpretation: if the firm perceives the expected legal standard of precaution to be sufficiently less than the social optimum, then the injurer will underprotect; conversely, if the potential injurer believes the legal standard to be significantly to the right of the optimum, it will overprotect. This interpretation highlights the importance, noted elsewhere, of setting the legal standard equal to the social optimum in order for negligence to induce efficient precaution (Cooter and Ulen, 1988).

III. Negligence and *Ex Ante* Regulation

We come now to the important public policy issue of whether efficiency is better served by joint use of a negligence rule and *ex ante* regulation (rather than negligence alone). We proceed by introducing an *ex ante* safety regulation into the model just developed.

Safety regulation typically specifies a minimally acceptable level of precaution. There is no uncertainty with regard to the regulatory constraint; that is, the firm and the regulatory agency know the level of the constraint[15] and that it is enforced with certainty.[16] Let the safety regulation specify that precaution must be at least s. How does information about the safety regulation influence the firm's perception about the (uncertain) legal standard of care? The firm knows that the legal standard of precaution cannot be less than s. But the firm may also perceive the legal standard to be significantly greater than s. This seems most clearly to approximate the prevailing relationship between *ex ante* and *ex post* regulation where they are jointly used. For example, no court today accepts compliance with a regulatory agency standard as a complete defense against a complaint of negligence.

We first examine the impact from introducing the safety regulation on the injurer's level of precaution in our previous model. We represent the introduction of *ex ante* regulation by changing the injurer's distribution around the legal standard of precaution. With a safety regulation, s, the injurer will not consider precaution below s. In effect, the firm's probability distribution on the legal standard is truncated at s. (By assumption, there is zero probability that the legal standard will be below s.) There are a number of assumptions that could be made about the firm's new truncated subjective distribution on the legal standard, $\underline{q}(x)$. And the reader should note that our remaining results hinge on the relationship between $q(x)$ and $\underline{q}(x)$.[17] We take a Bayesian approach, making a simple and direct assumption that $\underline{q}(x)$ has a conditional distribution

$$\underline{q}(x) = q(x|x \geq s). \tag{16}$$

In other words, the probability mass that did lie below s is now distributed above s. Thus, we can write the conditional probability $\underline{R}(x)$ that the injurer will pay damages if

[15] No uncertainty with respect to the regulation may seem inconsistent with the assumption of uncertainty for liability. However, the asymmetry is real. Regulations are promulgated *ex ante*. The legal standard of care on the other hand is rarely known with the same degree of precision.

[16] Certainty of enforcement implies that the sanction for noncompliance is sufficiently high that the firm never adopts a level of precaution below the minimally acceptable level that is the safety regulation.

[17] An alternative assumption would be that all of the probability mass, $1 - R(s)$, is deposited at s with $q(x) = q(x)$ for $x > s$. In this case Propositions 4 and 5 do not hold. Alternatively, as a referee has suggested, a regulation may decrease the probability that a court will find a higher legal standard. An example of this is the tobacco companies' successfully arguing that their compliance with rules on warning labels insulates them from tort liability.

its level of precaution is x as

$$\underline{R}(x)=\frac{R(x)}{R(s)}. \tag{17}$$

The objective function of the injurer who is subject to both a safety regulation and negligence liability becomes

$$\min_{x} \underline{TC}(x)=C(x)+A(x)\underline{R}(x). \tag{18}$$

Let $\hat{x}$ be the level of precaution that satisfies this minimization problem. Then $\hat{x}$ can be viewed as a function of $s, \hat{x}(s)$. For a given s, the first-order condition for $\hat{x}$ can be written as

$$\underline{TC}'(\hat{x})=R(s)C'(\hat{x})+A'(\hat{x})R(\hat{x}) - A(\hat{x})q(\hat{x})=0, \tag{19}$$

where $\hat{x}$ is understood to mean $\hat{x}(s)$.

The question that needs to be addressed is how the injurer's choice of $\hat{x}$ changes with a change of the *ex ante* safety regulation s; that is, what is the sign of $d\hat{x}/ds$? The answer to this question requires the total differentiation of the first-order condition given in (19). The result of this total differentiation, upon rearrangement of terms, is

$$\frac{d\hat{x}}{ds}=\frac{q(s)C'(\hat{x})}{\begin{array}{c}R(s)C''(\hat{x})+A''(\hat{x})R(\hat{x})\\ -2A'(\hat{x})q(\hat{x})-A(\hat{x})q'(\hat{x})\end{array}}. \tag{20}$$

By assumption, $C(x)+A(x)R(x)$ is convex. Thus, the denominator of (20) is positive. The numerator is also positive, which implies that $d\hat{x}/ds$ is greater than zero. Thus, increasing the minimally acceptable safety regulation has the effect of increasing the precaution taken.

The above result would imply that if $\bar{x} < x^*$ (that is, $\hat{x}(0) < x^*$) prior to the imposition of the *ex ante* regulation, then the introduction of the regulation will promote efficiency. If, on the contrary, $\hat{x}(0) > x^*$, then the *ex ante* regulation will exacerbate the inefficiency that exists with the negligence rule.

PROPOSITION 4: *Imposition of an ex ante regulation, given the existence of a negligence rule, will promote efficiency if the injurer were to be underprotective regarding an accident without ex ante regulation and will exacerbate inefficiency if the injurer were to employ too high a level of precaution without ex ante regulation.*

The gist of our result is that the imposition of an *ex ante* regulatory standard, on top of a negligence standard, induces potential injurers to revise their perceptions of the legal standard of care. Specifically, the higher the level of the *ex ante* regulatory standard, the higher the legal standard is likely to be, at least in the eyes of the injurer. Thus, the *ex ante* regulation can correct cases of underprecaution resulting from exposure to liability alone, but it can also exacerbate overprecaution.

This proposition may now be related to the conclusions of the previous section regarding the injurer's likely response to a negligence rule with uncertain enforcement of the legal standard. Recall that Propositions 1 through 3 established that injurers, when faced with only a negligence rule, may choose suboptimal precaution when (1) uncertainty about the legal standard is sufficiently large; (2) the marginal cost of precaution at x^* is large; or (3) the distribution about the legal standard is sufficiently biased to the left of x^*.

It follows that when any of these conditions holds, injurers can be induced to increase their level of precaution by establishing a minimum safety regulation, s. Additionally, it follows from our discussion that because $d\hat{x}/ds$ is always positive, the imposition of an *ex ante* minimum level of precaution in circumstances other than those noted above will cause injurers to take too much precaution.

Given that the introduction of an *ex ante* safety regulation can reduce inefficiencies associated with the use of liability alone, the obvious next question is what level of the *ex ante* regulation, s^*, will induce firms to choose $\hat{x}(s)=x^*$? From Proposition 4, we know that $s^*=0$ if and only if $\hat{x}(0)\geq x^*$. Furthermore, if $\hat{x}(0)<x^*$, then $s^*>0$ will promote efficiency. The question now is, what level of s will make $\hat{x}=x^*$? The answer can be found by substituting x^* for $\hat{x}$ in equation (19) and solving for s^*. Stopping short of actually solving equation (19) for s^*, we can rewrite it, using equation (2), as

$$C'(x^*)[R(s^*)-R(x^*)] - A(x^*)q(x^*)=0. \tag{21}$$

For this to hold, the bracketed term must be nonnegative. This implies that the optimum level of the *ex ante* regulation is less than or equal to the optimal level of precaution, that is, $s^*\leq x^*$. Furthermore, generally $s^*<x^*$. Consider the implications of $s^*=x^*$. In this case $R(x^*)=R(s^*)$, which implies, using equation (21), that the probability density at the social optimum ($q(x^*)$) must equal zero (because $A(x^*)>0$). In other words, the only way s^* can equal x^* is if there is no chance that the legal standard will be at x^*. That is unlikely.[18]

PROPOSITION 5: *The optimum level of an ex ante safety regulation*, s^*, *given that a negligence rule exists, will be less than the socially optimal level of precaution*, x^*, *provided* $q(x^*)>0$. *If* $q(x^*)=0$, *then* $s^*=x^*$ *is optimal.*

The implication of this result is that where optimal precaution calls for the joint use of *ex ante* regulation and a negligence rule, the optimal *ex ante* regulatory constraint should be set below the socially optimal level of care unless there is no uncertainty concerning the legal standard of care.

This result is illustrated in Figure 2 for two cases. Case I is the situation where negligence on its own overprovides precaution, x^+. An *ex ante* regulation cannot increase efficiency. In fact, for any $s>0$ in case I, the level of precaution, $x_1(s)$, increases and deviates even further from the social optimum. Case II involves underprovision of precaution when the firm is subject only to liability regulation, x^-. We see in both cases that as the *ex ante* regulation is raised, precaution increases. In case II, the optimal *ex ante* regulation is where s^* results in precaution of x^*. Also shown in the figure is what we might call the "conventional wisdom" along the kinked line: liability alone induces optimal behavior for $s<x^*$; as soon as s reaches x^*, then the *ex ante* regulation becomes binding and precaution is provided at level s.

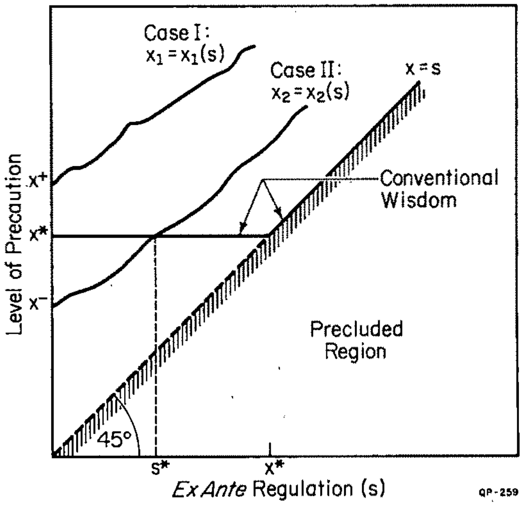


FIGURE 2. THE EFFECT OF *EX ANTE* REGULATION WITH *EX POST* LIABILITY

IV. Conclusions

The propositions presented above have profound implications in a wide range of public policies for dealing with external costs with regard to the conditions where *ex ante* regulation should be used alone or both *ex ante* regulations and *ex post* liability rules should be used jointly. We introduced uncertainty into a defendant's assessment of

[18] In (20), the term $C'(x^*)$ is positive as is the term $A(x^*)$, given our earlier definitions of these functions. In order for equation (20) to equal zero, $R(s^*)$ (the probability that s^* is less than the *ex post* standard) must be greater than that same probability at x^*. This relationship between $R(s^*)$ and $R(x^*)$ can only be true if $s^*<x^*$.

the legal standard of care and deduced the consequences of this uncertainty on the defendant's choice of precaution under a negligence rule.

Propositions 1 through 3 indicate the effect of uncertainty about the legal standard of care and the injurer's marginal cost of precaution on under- or over precaution. We next demonstrated that the introduction of an *ex ante* constraint specifying a minimally acceptable level of precaution (a safety regulation) will always cause the injurer to increase precautionary levels. We concluded that the joint use of *ex ante* and *ex post* regulation will enhance efficiency under the following conditions: if there is great uncertainty in the determination of the legal standard of care, or if the distribution is highly biased to the left of the socially optimal level of care, or if the injurer's marginal cost of precaution is large at the social optimum. Otherwise, *ex ante* and *ex post* regulation should be used separately.

We used our model to show the relationship between the optimal *ex ante* constraint and the socially optimal level of care. Proposition 5 clearly indicates that if it is efficient to use both policies, then the level of the *ex ante* regulation should not be set at the social optimum but rather at a lower level. That proposition might further be taken to indicate that *ex ante* regulations should be used alone when the probability of a successful suit against the injurer is zero. This might be the case when a great deal of uncertainty is associated with a harm, as might occur when the harm is so new that those it affects and the consequences of the harm are unclear but suspected of being catastrophic, or when the level of accident costs borne out by the injured party is so small that he or she might not even recognize it, even though many individuals are affected.

Further implications for the optimal mix of regulatory policies arise from a comparison of the assumptions of Propositions 1 through 4 with actual circumstances. This comparison may reveal both positive and normative insights. For example, it might be possible, using the model discussed here, to explain why new harms—for example, the escape of toxins into the environment—are typically regulated through *ex ante* command and control policies, while harms arising from more familiar sources—for example, automobiles—are typically regulated by exposing the injurer to *ex post* liability.

There are several refinements to the model that seem appropriate. First, our model includes only the injurer's costs of avoiding the harm. A complete analysis would include an explicit treatment of the victim's precautionary behavior under uncertainty. Second, a comparison of the administrative costs of the tort liability system and of the *ex ante* system should be made. Third, uncertainty surrounding the legal standard could be further broken down into its different components—for example, evidentiary uncertainty, uncertainty regarding the technology of precaution, uncertainty regarding the level of accident costs, and uncertainty about the victim's willingness to bring a tort action—in order to allow the examination of the conditions under which alternative *ex post* liability rules or a different mix of *ex ante* and *ex post* regulation might be efficient. Fourth, the possibility of bankruptcy could be introduced. Finally, uncertainty regarding the *ex ante* regulation could be introduced into the model. We have assumed that there is no error in the determination or enforcement of the *ex ante* standard. To the extent that such uncertainty exists, then the case for complementary use of *ex ante* and *ex post* regulation becomes more complex. While it is clear that such uncertainty would not affect our prohibition of the use of *ex ante* regulation when injurers tend to oversupply precaution because of uncertainty regarding the enforcement of the *ex post* liability rule, it should affect the level of regulation when injurers undersupply precaution. It could be argued that if regulators have enough information to set a lower bound on precaution *ex ante*, then the liability system should also have enough information to set the same lower bounds. But this reform cannot be affected within the tort liability system as it presently exists. It can only be achieved by supplementing exposure to tort liability with exposure to *ex ante* regulation.

REFERENCES

Baumol, William J. and Oates, Wallace E., "The Use of Standards and Prices of Protection of the Environment," *Swedish Journal of Economics*, March 1971, *73*, 42–54.

Brown, John P., "Toward an Economic Theory of Liability," *Journal of Legal Studies*, June 1973, *2*, 323–50.

Calabresi, Guido, *The Costs of Accidents: A Legal and Economic Analysis*, New Haven, CN: Yale University Press, 1970.

______ **and Klevorick, Alvin**, "Four Tests for Liability in Tort," *Journal of Legal Studies*, 1985, *14*, 585–628.

Calfee, John E. and Craswell, Richard, "Some Effects of Uncertainty on Compliance with Legal Standards," *Virginia Law Review*, 1984, *70*, 965–1003.

Cooter, Robert, Kornhauser, Lewis and Lane, David, "Liability Rules, Limited Information and the Role of Precedent," *Bell Journal of Economics*, Spring 1979, *10*, 366–73.

______ **and Ulen, Thomas S.**, "An Economic Case for Comparative Negligence," *New York University Law Review*, December 1986.

______, *Law and Economics*, Glenview, IL: Scott, Foresman & Co., 1988.

Craswell, Richard and Calfee, John E., "Deterrence and Uncertain Legal Standards," *Journal of Law, Economics and Organization*, 1986, *2*, 279–303.

Diamond, Peter, "Single Activity Accidents," *Journal of Legal Studies*, 1974, *3*, 107–64.

Ellikson, Robert, "Alternatives to Zoning: Covenants, Nuisance, and Fines as Land Use Controls," *University of Chicago Law Review*, Summer 1973, *40*, 681–781.

Lippman, Steven A. and McCall, John J, "The Economics of Uncertainty: Selected Topics and Probabilistic Methods," in K. Arrow and M. Intrilligator, eds., *Handbook of Mathematical Economics*, Amsterdam: North-Holland, 1981, 211–84.

Meyer, Jack, "Two-Movement Decision Models and Expected Utility Maximization," *American Economic Review*, June, 1987, *77*, 421–30.

Posner, Richard A., "The Jurisprudence of Skepticism," *Michigan Law Review*, 1988, *86*.

Roberts, Marc J. and Spence, Michael, "Effluent Charges and Licenses Under Uncertainty," *Journal of Public Economics*, April-May 1976, *5*, 193–208.

Rose-Ackerman, Susan, "Effluent Charges: A Critique," *Canadian Journal of Economics*, 1973, *6*, 512–28.

Shavell, Steven, (1984a) "Liability for Harm Versus Regulation of Safety," *Journal of Legal Studies*, June 1984, *13*, 357–74.

______, (1984b) "A Model of the Optimal Use of Liability and Safety Regulation," *Rand Journal of Economics*, Summer 1984, *15*, 271–80.

______, *Economic Analysis of Accident Law*, Cambridge, MA: Harvard University Press, 1987.

Ulen, Thomas S., Hester, Mark and Johnson, Gary V., "Minnesota's Environmental Response and Liability Act: An Economic Justification," *Environmental Law Reporter*, April 1985, *15*, 10109–115.

Weitzman, Martin L., "Prices Versus Quantities," *Review of Economic Studies*, October 1974, *41*, 477–91.

Wittman, Donald, "Prior Regulation Versus Post Liability: The Choice Between Input and Output Monitoring," *Journal of Legal Studies*, January 1977, *6*, 193–211.

The Subsidence of Preference Reversals in Simplified and Marketlike Experimental Settings: A Note

By YUN-PENG CHU AND RUEY-LING CHU*

A behavior pattern called the "preference reversal phenomenon" has received a lot of attention in recent years. This phenomenon refers to the experimental finding that, when individuals are presented with two gambles—one having a high probability of winning a modest sum of money (the *P* bet), and the other having a low probability of winning a large amount of money (the $ bet)—they often prefer the *P* bet but assign a larger monetary value to the $ bet. Such behavior is astonishing because, as Paul Slovic and Sarah Lichtenstein (1983) noted, "it violates almost all theories of preference, including expected utility theory."

Considerable efforts have been made in the literature to better understand this puzzle on both the theoretical and experimental fronts. So far theoretical efforts to rationalize such behavior appear to be quite fruitful; notable examples are Graham Loomes and Robert Sugden (1983), Chew Soo Hong (1983), Charles A. Holt (1986), Ronald A. Heiner (1983, 1985), Vernon L. Smith (1982, 1985), Paul Slovic and Sarah Lichtenstein (1983), William M. Goldstein and Hillel J. Einhorn (1987), Edi Karni and Zvi Safra (1987), and Uzi Segal (1988). The results from experimental studies have, however, been much less encouraging, in the sense that the behavior pattern remains robust in spite of the many different kinds of improvements in experimental design.[1] It is apparently still an open question whether this phenomenon will ever disappear or at least subside under still newer experimental settings. The current study is an effort to answer this question. It consists of two experiments. The features of these experiments and the reasons for adopting them are explained below.

(i) Most of the recently performed experiments have their subjects work with several pairs of gambles. In the experiments of Grether and Plott and Robert J. Reilly (1982), subjects deal with a total of twelve (six pairs of) gambles and go through three steps. First, they are shown three of the six pairs and asked to indicate a preference for each pair. Second, they are given all twelve gambles, presented in no particular arrangement, and asked to assign a monetary value to each. Finally, they are presented with three pairs of gambles, which are the ones left over from step one, and are required to indicate a preference for each pair. In Werner W. Pommerehne, Friedrich Schneider, and Peter Zweifel's (1982) experiment, which has a similar format, the total number of gambles is eight, arranged into four pairs in steps one and three. In Berg et al.'s experiment, subjects go through two sections, each of which duplicates the format of Grether and Plott (and Reilly) described above, with twelve gambles and three steps. One question naturally arises: Will the reversal phenomenon be reduced in less complex experiments? The current study examines this issue by confining the subjects' choice to a single pair of gambles in the main parts of the experiments.

(ii) As Smith (1985) correctly points out, a preference reverser is vulnerable to a con game (money pump) in which a sequence of decisions will cause a loss in assets. It would then be interesting to know whether the phenomenon would persist in a marketlike environment where reversers would see ac-

*Sun Yat-Sen Institute for Social Sciences and Philosophy (formerly Institute of the Three Principles of the People), Academia Sinica, Taipei 11529, Taiwan, Republic of China. Parts of this study were made possible by a grant from National Science Council, Executive Yuan, Republic of China (No. NSC74-0301-H001-06). The authors thank an anonymous referee for his immensely helpful comments and suggestions.

[1] See Joyce E. Berg, John W. Dickhaut, and John R. O'Brien (1985) and David M. Grether and Charles R. Plott (1979) for a good review of the experiments that have been performed.

tual losses in their assets as the result of arbitrage by the experimenter. Berg et al. perform precisely such an experiment in which they find that although the value (dollar magnitude) of reversals is reduced, the frequencies of reversals are not. One may, however, question whether their experiment is a serious disproof of the usefulness of a marketlike environment on two grounds: One is the problem of complexity as noted above. The other is that their experiment exposes the subjects to arbitrage only once, and it is questionable whether one exposure is enough to reduce reversals. The experiments in the current study are designed to deal with both problems. In particular, in the second part of our first experiment, subjects are presented with only two gambles. Each subject is asked to indicate a preference between, and the monetary values of, the gambles, and one arbitrage is undertaken if there is a reversal. The subject is then asked to state the preference and the monetary values again; arbitrage is again undertaken if there is reversal, and this process is repeated several times. In our second experiment, to see whether the subjects are able to carry their experiences from one pair of gambles to another, we perform three consecutive con games, each of which consists of a series of arbitrages. These experiments represent, we think, a fairer test of the robustness of the reversal phenomenon in the face of a marketlike environment.

(iii) Since the current study was carried out in Taiwan, while most of the other experiments so far performed took place in Western industrialized nations, it is necessary to ensure the cross-cultural applicability of the reversal phenomenon. The first part of our first experiment does this by completely replicating Reilly's process.

To summarize, the current study contains two experiments, with the first one further divided into two parts, I and II. Part I of the first experiment replicates Reilly's design, and Part II is a con game consisting of a series of arbitrages, played for a single pair of gambles. The second experiment goes one step further by performing three consecutive con games for three different pairs of gambles. In what follows, Section I will explain in detail how the two experiments were carried out; Section II will report the results; and, finally, Section III will offer some concluding remarks.

I. Experiment Design

A. *First Experiment*

There were two groups of subjects. One group consisted of 39 freshmen in applied psychology, and the other consisted of 55 juniors in economics, both at Fu-Jen Catholic University in Taiwan. They will be referred to as Group A and Group B, respectively.

The first part, or Part I, of our experiment essentially copies Reilly's Group 1 in Stage 2. As noted briefly earlier, it involved three steps. First, subjects were shown three pairs of gambles and asked to indicate their preference within each pair. Second, the subjects were shown twelve gambles, including the six from the previous step, unpaired and randomly presented. For each gamble, the subjects were asked to record the lowest price for which they would part with the right to play the gamble. Finally, subjects were shown three more pairs of gambles, made of the six gambles that appeared new in step two, and were asked to select the gamble they preferred in each pair.

Most of the details of the above steps can be found in Reilly, so they are not repeated here. We will only note those arrangements of ours that are not explained in detail in Reilly or are different from his arrangements, as follows:

(i) The whole group went through this part together in a classroom. They were not divided into small sections.

(ii) At the beginning of Part I, subjects in Group A were given a cash payment of NT$100 (US$3.50 at the average exchange rate of NT$28.59:US$1 in 1988) each for participating in the experiment, and those in Group B were paid NT$50 (US$1.75). The difference between the two sums was due to administrative reasons. Then each of the subjects, in groups A and B, was staked NT$90 (US$3.15) in token money that could be converted into cash at the end of the experiment. Given that Taiwan's per capita income is less than one-third of that of the United States, these amounts are certainly comparable to if not more attractive than those in Reilly's experiment.

(iii) The six pairs of gambles used in this Part I are (33/36; 70, −80) and (21/36; 140, −60); (33/36; 40, −40) and (18/36; 100, −30); (29/36; 60, −50) and (14/36; 210, −70); (31/36; 80, −30) and (12/36; 270, −40); (32/36; 60, −40) and (14/36; 170, −30); and (34/36; 60, −40) and (18/36; 130, −20); where the fractions indicate the probabilities of winning, and the winning and losing amounts are all in NT dollars. They will be called P1, P2,...,P6, respectively. These pairs are the same as Reilly's except for the payoffs, which we think are attractive enough, as far as our subjects are concerned, for the reasons stated above.

(iv) At the end of Part I, one of the six pairs of gambles was chosen by the random process of drawing numbered Ping-Pong balls from a jar. Each subject was assigned one of the gambles in that pair according to his stated preference. Then the experimenter disclosed to the subjects a "transaction price" for each of the two gambles—that is, a price that the experimenter could compare with the subject's recorded selling price (which could be negative) to determine whether he would buy the gamble from the subject or let the subject play it.[2] The subjects had already been told at the beginning of the game that this transaction price was an amount between NT$270 and −NT$30 (the maximum possible amounts of winning and losing for all gambles) that had been selected by a random process prior to the game. Now, if the gamble assigned to a subject should be bought, the experimenter bought it and paid the subject the purchasing price in token money. If the gamble should be played, it was played, and the appropriate sums were paid to or collected from the winners and losers in the form of token money. So at the end of Part I, each subject had either sold or played a gamble. Each was paid the actual value of his token money in cash at the end of the experiment.

After Part I was completed, all subjects in the group were asked to stay in the classroom and no conversation was allowed. Part II began in a separate classroom with a total of five experimenter's assistants ("traders") sitting in five separate parts of the room. Each trader "handled" only one subject at a time. No conversation was allowed between subjects in this room. Also, the subjects were not allowed to return to the original classroom, which was used as a waiting room, after the completion of Part I.

So Part II began as the experimenter randomly chose students from the waiting room and assigned each of them to one of the traders, also randomly. It took an average of one minute and a half for a trader to handle one subject. Once the process was completed, the subject was asked to leave the room, and a new subject was summoned from the waiting room. After all the subjects had completed their transactions with the traders, Part II was completed.

The sequence of interactions between the subject and the trader was arranged as follows.

1. The subject was given a new type of token money (colored differently from the money used in the first part) worth NT$300 (US$10.49) and was told, "This is the initial stake for this part of the game. It is not real money because it cannot be cashed. But whether you win or lose in this part of the game still matters, because at the end, I will give you a present, a real gamble with a large positive gain and a small loss, which you can choose to play or not to play. If you decide to play and win, I will pay you the same type of token money used in the last part, and it will be cashed. If you lose, I will take the losing amount from the token money you are keeping. So the gamble will be real. There are many gambles in my drawer, some of them are very attractive (with higher winning amounts and lower risks) and some are less attractive. Which of them I will give you as a present depends on

[2] This is what is usually called the Becker-DeGroot-Marschak procedure. See Reilly for details.

the total value of your new type of token money at the end of this part of the game. The better you do, the more valuable a present you will receive.[3]

2. The subject was asked to study two gambles, α and β, which are the pair of gambles that invited the highest frequency of (positive) reversals in Part I, for either Group A or Group B.[4] This pair turned out to be P4; that is, $\alpha=(31/36;\ 80,\ -30)$ and $\beta=(12/36;\ 270,\ -40)$.

3. The subject was asked to state his preference between α and β and was told that if later he were holding a gamble that was not the one he preferred, or if he had no preference between the two gambles, then the trader would have the right to exchange the other gamble for the one he was holding.

4. The subject was asked to indicate a "fair" price for each α and β and was told that whatever the price (positive or negative) might be, if later the trader decided to buy from or sell to him gamble α or gamble β at the indicated fair price,[5] he must comply.

The subject's response to steps 3 and 4 above must fall under one of the following categories:

(i) $\alpha \succ$ (is preferred to) β and $\$(\alpha)$ (the fair price of $\alpha) \geq \$(\beta)$, or $\alpha \prec \beta$ and $\$(\alpha) \leq \$(\beta)$, or $\alpha \sim$ (is as preferred to as) β and $\$(\alpha)=\$(\beta)$. In these cases no transaction took place; the subject was told that the game was over, his new token money (NT$300) was taken back, and he was given a gamble as a present. If the subject decided not to play this gamble, he was asked to leave; if he decided to play, the gamble was played, the winning or losing amount was settled, and he was asked to leave. Part II was completed.

(ii) $\alpha \succsim \beta$ and $\$(\alpha)<\$(\beta)$. The trader now sold β to the subject, exchanged α for β, and bought α back. The subject found that he was not holding any gambles but his (new type of) token money had decreased by $\$(\beta)-\$(\alpha)$. Then the subject was asked to go back to steps 3 and 4 above.

(iii) $\alpha \precsim \beta$ and $\$(\alpha)>\$(\beta)$. Analogously, one arbitrage transaction took place and the game went back to steps 3 and 4.[6]

For Group A, a maximum of two arbitrage transactions were undertaken for any one subject, no matter how many times he reached situation (ii) or (iii). For Group B, no maximum was set and arbitrage continued unless situation (i) was reached.[7] In either case, once no more arbitrage transaction was to take place, the game was over and the subject went through the same procedure as that in situation (i). This completed Part II.

On the next day, the token money held by the subjects was cashed, and the experiment was completed.

B. *Second Experiment*

Again, there were two groups of subjects; they will be referred to as groups C and D. Group C consisted of 46 freshmen in Chinese literature at Fu-Jen Catholic University, and Group D consisted of 37 freshmen with the same major at National Cheng-Chi University.

[3] Admittedly, the use of an unknown gamble as a reward instead of real money constitutes a deficiency in our experimental design, as pointed out to us by an anonymous referee. However, this shortcoming is remedied in the second experiment, which is simply a more complicated version of Part II of the first experiment.

[4] The meaning of "positive" reversals will be explained later.

[5] If the price is negative (positive), the trader will take back from (give to) the subject some token money in order to obtain a game from the subject. In reality none of the subjects stated a negative price for any gambles in this part of the game or in the second experiment (to be discussed later).

[6] If in the process of arbitrage, the value of the new type of token money held by the subject fell below zero, he would be floated a loan. This situation did not, however, happen in reality.

[7] For Group A, which was given the experiment first, we had no idea how rapidly or slowly subjects would change their reversal behavior under arbitrage, so we set the maximum at two to control time. No maximum was set for Group B because we knew from our experience with Group A that reversers changed their behavior rapidly. See the later text for details.

TABLE 1—FREQUENCIES OF REVERSALS, ALL SIX PAIRS

			Reservation Prices (Percent)[b]			
	Bet	Choices[a]	Consistent	Inconsistent	Equal	Total[c]
Group A	*P*	110	45.5	47.3	7.3	100.0
$n = 39$	$	124	67.7	31.5	0.8	100.0
Group B	*P*	132	45.5	48.5	6.1	100.0
$n = 55$	$	195	70.8	23.6	5.6	100.0
Groups A and B	*P*	242	45.5	47.9	6.6	100.0
$n = 94$	$	319	69.6	26.6	3.8	100.0
Reilly's Results[d]	*P*	106	46.2	49.1	4.7	100.0
$n = 45$	$	159	67.9	26.4	5.7	100.0

[a] The total number of choices may not be equal to the total number of possible choices, namely $6 \times n$, because some subjects are indifferent between the P and $ bet.

[b] Frequencies of consistent, inconsistent or equal prices divided by the total number of choices in the same row.

[c] Actual sum may not be exactly 100.0 because of rounding.

[d] This is his Group 1 in Stage 2.

This experiment skipped Part I of our first experiment and went directly to Part II, with the following added features:

(i) The gambles used in this experiment were the three pairs that invited the highest numbers of positive reversals in Part I of our first experiment from groups A and B combined. It is already known that P4 is one, and the other two pairs turned out to be P1 and P5.

(ii) Each subject was first given NT$50 (US$1.75) in cash for attendance. The experimenters then spent about 15 minutes explaining the gambles and the rules of transactions, which were the same as those in Part II of the first experiment.

(iii) Three different sequences of con games were designed, they were, respectively, P1-P4-P5, P4-P5-P1 and P5-P1-P4. Each subject was randomly assigned to one of the three sequences.

(iv) Using P1-P4-P5 as an example: a subject was first staked NT$300 (US$10.49) in token money. The trader then played P1 with him, following the same steps as those in Part II of the first experiment, without any limit on the maximum number of arbitrage transactions. When these steps were completed, the subject kept his token money, however much it was worth at that point, and was again given NT$300 (US$10.49) in token money. Pair P4 was then played. This process was repeated until pair P5 was also played. The subject was told that the total amount of token money he had at the end of the three con games would be cashed at 2/9 of its face value. Thus if a subject lost no money in these con games, he would receive NT$200 (US$7.00) in the end, in addition to the NT$50 (US$1.75) attendance reward.

This completes the description of our second experiment.[8]

II. Results of the Experiments

Table 1 gives our Part I results and those of Reilly. It is clear that our subjects in Group A or Group B or both together were remarkably similar to Reilly's in their preference reversal behavior. The occurrence of "positive" reversals (P bet is preferred but assigned a lower monetary value) was 47.3 percent for Group A, 48.5 percent for Group B, and 47.9 percent for both groups, while it was 49.1 percent for Reilly's Group 1 in his Stage 2. As for the "negative" reversals ($ bet is preferred but assigned a lower value), the rate was 31.5 percent for our Group A, 23.6 percent for Group B, and 26.6 for both groups, while the rate for Reilly's group was

[8] The whole idea of this second experiment was brought to our attention by an anonymous referee, to whom we are grateful. This explains why the two experiments in this paper could have been combined but were not.

TABLE 2—PREFERENCE REVERSALS UNDER ARBITRAGE: ONE PAIR OF GAMBLES

Group A: A Maximum of Two Arbitrage Transactions ($n = 39$)				
	Situation (i) $\alpha \succ \beta$ and $\$(\alpha) \geq \$(\beta)$ $\alpha \prec \beta$ and $\$(\alpha) \leq \$(\beta)$ or $\alpha \sim \beta$ and $\$(\alpha) = \$(\beta)$	Situation (ii) $\alpha \succsim \beta$ and $\$(\alpha) < \$(\beta)$	Situation (iii) $\alpha \precsim \beta$ and $\$(\alpha) > \$(\beta)$	Sum
Before Arbitrage	22[a] (56.4)[b]	16 (41.0)	1 (2.6)	39 (100.0)
After 1 Arbitrage Transaction	12 (30.8)	5 (12.8)	0 (0.0)	17 (43.6)
After 2 Arbitrage Transactions	3 (7.7)	2 (5.1)	0 (0.0)	5 (12.8)

Group B: No Limit on Number of Arbitrage Transactions ($n = 55$)				
	Situation (i) $\alpha \succ \beta$ and $\$(\alpha) \geq \$(\beta)$ $\alpha \prec \beta$ and $\$(\alpha) \leq \$(\beta)$ or $\alpha \sim \beta$ and $\$(\alpha) = \$(\beta)$	Situation (ii) $\alpha \succsim \beta$ and $\$(\alpha) < \$(\beta)$	Situation (iii) $\alpha \precsim \beta$ and $\$(\alpha) > \$(\beta)$	Sum
Before Arbitrage	43 (78.2)	11 (20.0)	1 (1.8)	55 (100.0)
After 1 Arbitrage Transaction	3 (5.4)	9 (16.4)	0 (0.0)	12 (21.8)
After 2 Arbitrage Transactions	8 (14.6)	1 (1.8)	0 (0.0)	9 (16.4)
After 3 Arbitrage Transactions	1 (1.8)	0 (0.0)	0 (0.0)	1 (1.8)

[a] Number of people.
[b] Percentage of total number of people in the group.

26.4 percent. These results indicate that our subjects are certainly comparable to Reilly's with regard to their preference reversal behavior, so our Part II results should be meaningful.

Data on the subjects' preferences between individual pairs of gambles are not given here but are available upon request. Suffice it to say that P4 invited the highest frequency of positive and negative reversals combined, with 35 positive and 9 negative ones; P1 ranked second with 25 positive and 13 negative reversals; and P5 ranked third with 18 positive and 17 negative reversals, for groups A and B combined. In addition, as stated earlier, when only the frequencies of positive reversals are counted, these same pairs also invited the highest numbers of reversals.

For Part II, the results of which are reported in Table 2, two of the most important findings are as follows:

(i) The frequency of reversals before arbitrage took place was reduced among the Group B subjects, but it remained exactly the same for Group A subjects. In Part I, 17 out of the 39 people in Group A made inconsistent choices with regard to P4. In Table 2, the numbers in the corresponding columns, namely situations (ii) and (iii) before arbitrage took place, add up to the same amount. For Group B, while in Part I the number of reversals is 27 out of 55 people for P4, it is reduced to 12 in Table 2. These results indicate that in the absence of arbitrage, preference reversal remained an important phenomenon even when the experiment was so simplified as to consist of a

TABLE 3—PREFERENCE REVERSALS UNDER ARBITRAGE: THREE PAIRS OF GAMBLES

	Round 1		Round 2		Round 3	
Group C ($n=46$)	People	Transactions[a]	People	Transactions[a]	People	Transactions[a]
Positive Reversal First Appears at the Beginning of						
Round 1	18	37	1[b]	1	0[b]	0
Round 2			7	11	1	3
Round 3					2	2
Negative Reversal First Appears at the Beginning of						
Round 1	2	2	1	1	0	0
Round 2			1	2	0	0
Round 3					2	2

	Round 1		Round 2		Round 3	
Group D ($n=37$)	People	Transactions[a]	People	Transactions[a]	People	Transactions[a]
Positive Reversal First Appears at the Beginning of						
Round 1	6	8	2	7	0	0
Round 2			1	3	0	0
Round 3					0	0
Negative Reversal First Appears at the Beginning of						
Round 1	3	3	1	2	0	0
Round 2			2	2	0	0
Round 3					2	3

	Round 1		Round 2		Round 3	
Groups C and D Combined ($n=83$)	People	Transactions[a]	People	Transactions[a]	People	Transactions[a]
Positive Reversal First Appears at the Beginning of						
Round 1	24	45	3	8	0	0
Round 2			8	14	1	3
Round 3					2	2
Negative Reversal First Appears at the Beginning of						
Round 1	5	5	2	3	0	0
Round 2			3	4	0	0
Round 3					4	5

[a] Total number of arbitrage transactions performed to convert all of the reversers (people) at the beginning of the indicated round to nonreversers.

[b] These people are subsets of the 10 people who are positive reversers at the beginning of Round 1, reported in the same row. The numbers of people in all other rows are defined analogously. In the next row for example, the 1 person under "Round 3" is a subset of the 7 people under "Round 2."

single pair of gambles, although simplification was effective in reducing reversal frequency for one group of the subjects.

(ii) The frequency of reversals decreased rapidly under arbitrage for both groups. From Table 2, for Group A, 12 out of the 17 people who were reversers before arbitrage adjusted their behavior to become nonreversers after one arbitrage transaction. Out of the remaining 5 people, 3 adjusted their behavior toward consistency after the second arbitrage transaction; and only 2 remained unconverted. Reversers in Group B, for which no limit was set on the number of arbitrage transactions, adjusted their behavior equally rapidly. Out of the 12 reversers before arbitrage, 3 became nonreversers after one transaction; out of the remaining 9, 8 became consistent players after the second transaction; and the remaining 1 person was also converted following the third arbitrage transaction. So for Group B, *three transactions were all that was needed to wipe out preference reversals completely.*

The results from our second experiment are presented in Table 3. In this table the three con game sequences referred to earlier—namely, P1-P4-P5, P4-P5-P1 and P5-P1-P4—have been combined. These three con games in any sequence are referred to as rounds 1, 2, and 3, respectively. That is, the first con game any subject encounters is called round 1, no matter to which of the three sequences the subject belongs; the second con game the same subject encounters is called Round 2, and so on.

What does one read from Table 3? Beginning with the positive reversers in Group C, one observes from the first row in Table 3 that of the 46 subjects in Group C, 18 displayed positive reversal behavior in the first round before arbitrage took place, and it took a total of 37 arbitrage transactions to convert all of them to nonreversers in this Round 1, an average of 2.06 transactions per subject. Then, given what the subjects learned in Round 1, only 1 out of these 18 people again displayed reversal behavior in the second round before arbitrage took place; and one arbitrage transaction was all that was needed to convert this person to a nonreverser in this round. Finally, none of these 18 people displayed reversal behavior anymore in Round 3, before arbitrage took place.

The second row in Table 3 shows that, 7 out of the 46 subjects in Group C displayed no reversal behavior in Round 1, but were positive reversers at the beginning of Round 2, before arbitrage took place. In this round, following a total of 11 arbitrage transactions, all 7 became nonreversers. Then, out of the 7 people who had been exposed to the arbitrage environment, only 1 again displayed reversal behavior in Round 3, and it took 3 arbitrage transactions to convert him to a nonreverser in that round.

Finally, according to the third row in Table 3, 2 out of the 46 people in Group C were nonreversers in rounds 1 and 2 but displayed positive reversal behavior in Round 3, before arbitrage took place. It took 2 arbitrage transactions to wipe out the behavior between these two subjects in that round.

Similar patterns were observed for the negative reversers in Group C and for the positive and negative reversers in Group D in Table 3 and need not be repeated. The overall picture Table 3 presents can be summarized as follows.

(i) For groups C and D combined, 29 (24 + 5) out of the 83 subjects reversed at the beginning of Round 1, and a total of 46 (24 + 8 + 2 + 5 + 3 + 4), or more than half of them, reversed at the beginning of one of the three rounds, prior to exposure to the money pump. These results indicate that in the absence of arbitrage, preference reversal remained an important phenomenon even when subjects worked with only a single pair of gambles at a time. Such a finding echoes that of Part II of the first experiment.[9]

[9] It is worth noting here that while it is meaningful to make a direct comparison between the results of the two parts of the first experiment to see whether simplification in design reduced reversal frequency, a similar direct comparison between the results of Part I of the first experiment and those obtained here in the second experiment would not be justifiable because subjects in this second experiment differed from those in the first

(ii) The within-round behavior of the reversers in Table 3 is similar to that of the subjects in Part II of the first experiment, as both groups of subjects were quick to adjust their behavior from that of reversers to nonreversers following a series of arbitrage transactions designed to make reversers lose money. The reversal phenomenon is vulnerable to arbitrage transactions.

(iii) For the 29 (24 + 5) subjects who first reversed at the beginning of Round 1 and for the 11 (8 + 3) subjects who first reversed at the beginning of Round 2, the between-round behavior clearly indicates that both sets of subjects were able to carry their experiences of arbitrage transactions from one round to subsequent rounds. The reversal phenomenon is vulnerable to the experiences the subjects accumulated from the previous rounds of marketlike games.

III. Concluding Remarks

The experiments reported in this paper are designed mainly to test three hypotheses: (i) incidence of reversals will be reduced if the choices are simplified; (ii) incidence of reversals will be reduced in a marketlike environment where reversers are subject to repeated arbitrage transactions that cause them to lose money; and (iii) incidence of reversals will be reduced for subjects who have had the experience of a marketlike environment.

From a comparison between the results of parts I and II of the first experiment, hypothesis (i) above is only weakly confirmed at best. Although reversal frequency was reduced for one group of subjects, it stayed at exactly the same level for the other group. Furthermore, an examination of the absolute numbers of reversal frequencies in Part II of the first experiment and in the second experiment reveals that preference reversal remained an important phenomenon, even when the choices confronting the subjects were reduced to a single pair of gambles. Simplification alone did not seem to be able to put the reversal phenomenon to rest.

But what simplification alone could not achieve became achievable once arbitrage was introduced. In Part II of the first experiment, only 5.1 percent of the subjects in Group A remained reversers after two transactions, and no one in Group B reversed after three transactions. In the second experiment, all reversers (in all rounds) were successfully converted to nonreversers after an average of 1.71 transactions per subject.[10] So hypothesis (ii) above is strongly confirmed.

Hypothesis (iii) is also strongly confirmed. Results of the second experiment show that subjects displayed substantially fewer reversals after they were exposed to a marketlike environment in previous rounds of games. They were able to carry their experiences from one game to subsequent games.

It seems, then, that while the preference reversal phenomenon is important, it is vulnerable to a marketlike environment coupled with a simplification in the experimental design. This is no doubt the most important finding of the current study.

Of course, the study as it is still leaves many important questions unanswered. We do not know to what extent or how long the "educational" effect of the con games will last. Will the subjects remember the "lessons" after six months or a year? Will they be able to apply their experiences to games that are of the same nature as the ones played here but have very different formats? Furthermore, in real life situations, if the payoffs at stake are very large, will they actively seek tests of their decision-making procedure in marketlike environments before they actually enter the markets, or they will still react to information transmitted to them from the market passively? That is,

in that they were not exposed to the Reilly's procedure (Part I of the first experiment) prior to making choices between a single pair of gambles.

[10] This average number of transactions is obtained by adding all of the numbers of transactions in Table 3 together, and dividing the sum by the total number of reversers in all categories, for groups C and D combined.

will they take precautionary measures to prevent mistakes from being made, or they will adjust their behavior only after the mistakes have already been made? A large body of literature already exists on these issues, but more experimental findings would still be helpful in scrutinizing the relative usefulness of alternative hypotheses. We look forward to future efforts in these directions.

REFERENCES

Berg, Joyce E., Dickhaut, John W. and O'Brien, John R., "Preference Reversal and Arbitrage," in V. Smith, ed., *Research in Experimental Economics*, Vol. 3, Greenwich, CT: JAI Press, 1985, 31–72.

Goldstein, William M. and Einhorn, Hillel J., "Expression Theory and the Preference Reversal Phenomena," *Psychological Review*, April 1987, *94*, 236–54.

Grether, David M. and Plott, Charles R., "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, September 1979, *69*, 623–38.

Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, *73*, 560–95.

______, "Experimental Economics: Comment," *American Economic Review*, March 1985, *75*, 260–63.

Holt, Charles A., "Preference Reversals and the Independence Axiom," *American Economic Review*, June 1986, *76*, 508–15.

Hong, Chew Soo, "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox," *Econometrica*, July 1983, *51*, 1065–92.

Karni, Edi and Safra, Zvi, "'Preference Reversal' and the Observability of Preferences by Experimental Methods," *Econometrica*, May 1987, *55*, 675–85.

Loomes, Graham and Sugden, Robert, "A Rationale for Preference Reversal," *American Economic Review*, June 1983, *73*, 428–32.

Pommerehne, Werner W., Schneider, Friedrich and Zweifel, Peter, "Economic Theory of Choice and the Preference Reversal Phenomenon: A Reexamination," *American Economic Review*, June 1982, *72*, 576–84.

Reilly, Robert J, "Preference Reversal: Further Evidence and Some Suggested Modifications in Experimental Design," *American Economic Review*, June 1982, *72*, 576–84.

Segal, Uzi, "Does the Preference Reversal Phenomenon Necessarily Contradict the Independence Axiom?" *American Economic Review*, March 1988, *78*, 233–36.

Slovic, Paul and Lichtenstein, Sarah, "Preference Reversals: A Broader Perspective," *American Economic Review*, September 1983, *73*, 596–605.

Smith, Vernon L., "Microeconomic Systems as an Experimental Science," *American Economic Review*, December 1982, *72*, 923–55.

______, "Experimental Economics: Reply," *American Economic Review*, March 1985, *75*, 265–72.

Tests of "Fanning Out" of Indifference Curves: Results from Animal and Human Experiments

By JOHN H. KAGEL, DON N. MACDONALD, AND RAYMOND C. BATTALIO*

In an earlier paper (Raymond C. Battalio, John H. Kagel, and Don N. Mac Donald, 1985), we reported Allais-type violations of the independence axiom of expected utility theory with rats choosing over positively valued payoffs (food rewards). This note extends this research, examining animals' choices over losses, testing for (1) standard Allais-type common ratio effect violations of expected utility theory and (2) fanning out of indifference curves for random prospects, tests of Mark J. Machina's (1982, 1987) hypothesis II (hereafter H2), over previously unexplored areas of the unit probability triangle. Results from a parallel series of experiments using human subjects choosing over real losses are also reported. For both rats and people, we find standard Allais-type violations of expected utility theory and a systematic failure of the fanning out hypothesis in the southeast corner of the unit probability triangle, in the case of losses. Thus, the fanning out hypothesis (Machina 1982, 1987) cannot provide a satisfactory explanation for behavioral deviations from expected utility theory.

I. Allais-Type Common Ratio Effects and New Tests of Fanning Out

A key element of all descriptive alternatives to expected utility theory is to explain Allais-type common ratio effects and a number of other systematic violations of expected utility theory. Common ratio effect violations of expected utility theory include the "certainty effect" of Daniel Kahneman and Amos Tversky (1979) and the "Bergen Paradox" of Ole Hagen (1979), as special cases. Allais-type common ratio violations involve rankings over pairs of prospects of the form:

Example 1

A: x_2 with probability p
  x_3 with probability $1-p$
B: x_1 with probability q
  x_3 with probability $1-q$

Example 2

C: x_2 with probability rp
  x_3 with probability $1-rp$
D: x_1 with probability rq
  x_3 with probability $1-rq$

where $p > q$, $0 \geq x_3 > x_2 > x_1$, and $0 < r < 1$ (the term "common ratio" derives from the equality of $\text{prob}(x_2)/\text{prob}(x_1)$ in A vs. B and C vs. D). An expected utility maximizer would prefer either A and C (if $p[u(x_2)-u(x_3)] > q[u(x_1)-u(x_3)]$), or else B and D (if the opposite were true). However, using primarily hypothetical choice alternatives, researchers find a systematic tendency for subjects to prefer B and C when the outcomes involve losses (Kahneman and Tversky, 1979; Kenneth R. MacCrimmon and Stig Larson, 1979; and the references cited in Machina, 1983).

Machina (1982, 1987) identifies three systematic violations of the independence axiom in addition to the common ratio effect and shows that all four follow from a single

*The authors' affiliations are, respectively, Department of Economics, University of Pittsburgh; Department of Finance, University of North Texas; Department of Economics, Texas A&M University. Research support from the Economics Division of the National Science Foundation is gratefully acknowledged. The research reported has benefited from discussions with, and the comments of, Colin Camerer, Mark Machina, and Amos Tversky and the comments of the referees. We alone are responsible for any errors.

assumption, H2.[1] More recently, Zvi Safra, Uzi Segal, and Avia Spivak (1988) have used H2 to explain the preference reversal phenomena (Sarah Lichtenstein and Paul Slovic, 1971; David W. Grether and Charles R. Plott, 1979; Machina, 1987). H2 states that in moving from one probability distribution to another, which (first order) stochastically dominates it, the local (linear in the probabilities) utility function retains the same degree of concavity, or becomes more concave at each point. In other words, preferences tend to vary systematically so that a first-order stochastic dominating shift in wealth (i.e., a nonnegative random addition to wealth) results in the same or more risk averse choices.

In terms of the unit probability triangle, this yields indifference curves that are parallel (as expected utility theory predicts) or that fan out relative to the origin. Figure 1 illustrates how fanning out can explain a typical Allais-type common ratio effect over losses. The set of all prospects over the fixed outcome levels $0 \geq x_3 > x_2 > x_1$ can be represented by the set of all probability triples of the form $P = (p_1, p_2, p_3)$ where $p_i = \text{prob}(x_i)$ and $\Sigma p_i = 1$. Since $p_2 = 1 - p_1 - p_3$, we can represent these lotteries by points in the unit triangle in the (p_1, p_3) plane. Figure 1 illustrates this for examples 1 and 2 above when $p = 1.0$, $q = 0.75$, and $r = 1/3$, (so that prospect A is represented by $(p_1, p_3) = (0,0)$, prospect B is represented by $(p_1, p_3) = (0.75, 0.25)$, etc.). Prospects C and D stochastically dominate A and B, as they are derived from A and B by reducing the probability of x_1 and x_2 by a common proportion, $0 < r < 1$, and adding the displaced probability mass to x_3 in both cases.

In Figure 1, the dashed lines represent linear combinations of the prospects chosen over. Solid lines are indifference curves over the random prospects. Since, in expected utility theory, the individual's indifference curves in the unit probability triangle are given by solutions to the linear equation

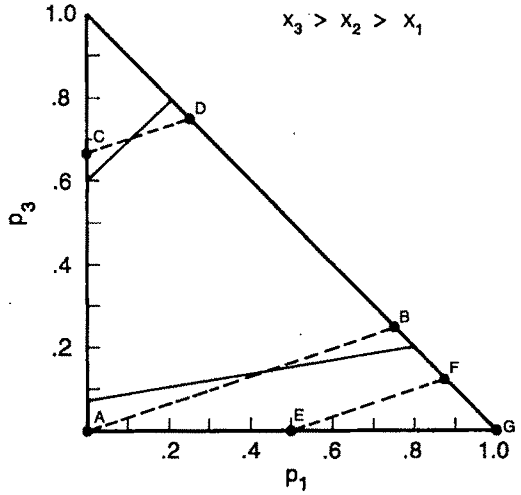


FIGURE 1. PROSPECTS EMPLOYED IN TESTS OF FANNING OUT. DASHED LINES CONNECT PROSPECTS' SUBJECTS CHOSE BETWEEN. STRAIGHT LINES ARE INDIFFERENCE CURVES THAT DISPLAY "FANNING OUT." PROSPECTS CORRESPOND TO THOSE SPECIFIED IN THE TEXT AND TABLE 1

$$\begin{aligned} \bar{U} &= \Sigma u(x_i) p_i \\ &= u(x_1)p_1 + u(x_2)(1 - p_1 - p_3) \\ &\quad + u(x_3)p_3 = \text{constant}, \end{aligned}$$

they will consist of parallel straight lines with slope $[u(x_2) - u(x_1)]/[u(x_3) - u(x_2)]$, with more preferred curves lying to the northwest. Since, by construction, the dashed lines are parallel, the parallel linear indifference curves of expected utility theory require preference for prospects A and C or B and D. However, indifference curves that "fan out" relative to the origin can, as shown in Figure 1, explain the preference for prospects B and C.

Fanning out of indifference curves in the unit probability triangle corresponds to the "light hypothesis" of weighted utility theory (Soo Hong Chew and Kenneth R. MacCrimmon, 1979) and is a characteristic of

[1] These involve (1) the common consequence effect, the most famous specific example of which is the so called "Allais Paradox," (2) over sensitivity to changes in small probability-outlying events, and (3) the utility evaluation effect (see Machina, 1982, 1987).

TABLE 1—TREATMENT CONDITIONS FOR LOSSES: ANIMALS

Condition	Prospects		
1	A: 5 sec delay, prob. 1.0 (5.0)	or	B: 13 sec delay, prob. 3/4 1 sec delay, prob. 1/4 (10.0)
2	C: 5 sec delay, prob. 1/3 1 sec delay, prob. 2/3 (2.34)	or	D: 13 sec delay, prob. 1/4 1 sec delay, prob. 3/4 (4.0)
3	E: 13 sec delay, prob. 1/2 5 sec delay, prob. 1/2 (9.0)	or	F: 13 sec delay, prob. 7/8 1 sec delay, prob. 1/8 (11.5)
4	G: 13 sec delay, prob. 1.0 (13.0)	or	F: 13 sec delay, prob. 7/8 1 sec delay, prob. 1/8 (11.5)
	Sequencing of Conditions: 1,2,1,3,1,3,1,4		

several other generalizations of expected utility theory as well (see Machina, 1983). In the case of losses, the obvious question is whether choices over pairs of prospects that are dominated by A and B, prospects, which lie in the southeast corner of the unit probability triangle and are parallel to A and B, such as E and F in Figure 1, satisfy the requirements of fanning out (preference for F on the part of those choosing B and C).

II. Experiment 1: Rats' Choices Over Losses

We define losses generically as any outcome for which *less* is better. In our experiments, the random variable was the time delay to reinforcement. Given the relatively steep temporal discount functions that rats have, time delay is a potent reinforcer, where it is well established that less is better (see Leonard Green, 1982, and Kagel and Green, 1987, for reviews of experiments most relevant to economists).

Historically, psychologists have studied choices between constant and variable delays to reinforcement, finding strong preference for the variable outcome alternative (risk loving), given equal average delays to reinforcement (Peter Killeen, 1968; see James E. Mazur, 1986, for a review of this literature).[2] In these experiments, there are no auxiliary decisions to be made between the choice point and the time that all uncertainty is resolved. This is important, since it is well established that in choosing over risky prospects whose resolution is not immediate, and where auxiliary decisions are required prior to the resolution of the uncertainty, the independence axiom of expected utility theory will not generally hold (Harry M. Markowitz, 1959, ch. 11; Jan Mossin, 1969).

Table 1 shows the variable outcome alternatives the rats chose between. Each outcome provided 0.1cc of water at the end of the delay period. Mean time delay to reinforcement is shown under each pair of prospects. Conditions 1 and 2 involve standard Allais-type common ratio manipulations as prospects C and D stochastically dominate A and B. Condition 3 checks for fanning out in the southeast corner of the unit probability triangle as prospects A and B stochastically dominate E and F and the dashed lines connecting the two sets of prospects are parallel. Prospects E and F are derived from A and B by multiplying

[2] Risk loving over losses here appears to be characteristic of a general preference for variable as compared to certain losses, consistent with predictions from prospect theory (Kahneman and Tversky, 1979). See Kagel et al. (1988, experiment 1), where we test this prediction by examining rats' preferences over variable versus certain shock with the same mean. The rats prefer the variable shock. Note that unlike humans choosing over monetary losses, there is no opportunity for "asset integration" over these choices, with its implication for inducing risk aversion.

the 5 and 1 sec delay probabilities by 0.5 and adding the displaced probability to the 13 sec delay probability in both cases. Condition 4 is a control condition that tests for the rats' sensitivity to low probability outcome events, since it is generally assumed that the decision maker knows the relevant probabilities of the different payoffs. Payoff frequencies of 1/8 may be sufficiently low that the rats are not cognizant of them. Condition 4 tests for this. Figure 1 shows these prospects in terms of the unit probability triangle.

A discrete trials choice procedure was employed where the rats chose between a single pair of prospects at a time. Choices were recorded and payoffs delivered in response to a single press on one of two choice levers. Each choice trial lasted 39 sec, irrespective of the alternative chosen or the length of the delay to reinforcement. This established a minimum of a 20 sec blackout period between choice trials (a maximum delay to reinforcement of 13 sec, with 6 sec to consume the water). Given the steep time discount rates the rats have, these 20 sec blackout periods create relatively strong independence between trials.

Each experimental session began with 32 forced choice trials where only one lever was operative, followed by 48 free choice trials when both levers were operative. The forced choice trials served to familiarize subjects with the alternatives. During these trials, the empirical distribution function was forced to match the programmed distribution function over each prospect's trial set.[3] Free choice trials served to measure preferences. The random number generator was given free reign during these trials, so that the probability of obtaining a given outcome on any trial was independent of outcomes on other trials.

Experimental sessions were conducted once a day, 7 days a week, at approximately the same time each day. Each experimental condition lasted 38 days, divided into two 19-day, periods. The levers delivering the certain and variable delay alternatives remained constant within each 19-day period, but they were switched between the two 19-day periods. Switching alternatives across levers and averaging choices controls for lever bias, which may be severe at times for rats. Preferences were measured in terms of choices of the certain alternative during the free choice trials for the last 5 days of each 19-day period and averaged across side switches.

Table 2 shows average percentage of free choice trials allocated to the more certain alternative under each condition. A choice frequency of 50 percent indicates that the rats chose the more certain alternative half the time, averaged across side switches, and is interpreted as indifference.

Under baseline pair 1 choices, the rats generally preferred the more variable (B) alternative. On average, the rats chose the A alternative on 38.9 percent of the trials ($t = -2.68$, $p < 0.05$, 2-tailed t-test). Using choice frequencies to measure intensity of preference, six of the eight subjects had relatively strong preference for the B alternative, the two exceptions being subjects 612 and 622. Given the strong preference for variable over fixed delay probabilities reported in the psychology literature, the preference for the B alternative here is not surprising, even though the average delay to reinforcement was twice that of the certain delay probability. Further, a necessary condition for generating Allais-type common ratio effect violations of expected utility theory, similar to those reported with human subjects, is to establish preference for the variable alternative under the high probability loss condition. The parameter values employed were designed to achieve this.

Under pair 2 choices, mean percentage choice for the more certain alternative, C, was 54.5 percent ($t = 1.63$). As indicated by the individual subject t-statistics, five rats were indifferent between C and D, and three had a clear preference for the C alternative. Upon return to pair 1 choices, mean percentage choice for the A alternative fell to 44.8 percent ($t = -2.01$, $p < 0.10$, 2-tailed

[3] Sequences of choices across levers and outcomes were fixed; however, the start point was determined randomly on a daily basis. Details of experimental procedures are reported in Kagel et al., 1988, experiment 2.

TABLE 2—CHOICES OVER LOSSES: ANIMAL SUBJECTS

	Percentage Choice of More Certain Alternative (Absolute Value of t-Statistics in Parentheses)							
		Standard		Test of		Test of		Sensitivity
	Baseline	Allais	Baseline	H2	Baseline	H2	Baseline	Test
Subj.	(1)	(2)	(1)	(3)	(1)	(3)	(1)	(4)
611	21.7 (7.26)[b]	72.3 (6.18)[b]	38.9 (3.58)[b]	64.6 (3.40)[b]	40.0 (5.18)[b]	64.4 (9.30)[b]	48.3 (0.60)	31.5 (5.56)[b]
612	50.0 (0.0)	50.2 (0.14)	55.6 (3.30)[a]	49.8 (0.54)	49.2 (1.00)	49.8 (0.28)	49.4 (1.34)	27.7 (10.2)[b]
613	44.6 (1.94)	49.6 (0.18)	50.4 (0.16)	67.7 (14.5)[b]	49.0 (0.34)	59.0 (4.92)[b]	46.0 (2.12)	31.5 (5.04)[b]
614	20.6 (15.1)[b]	48.1 (0.74)	32.5 (5.00)[b]	44.6 (1.68)	47.5 (1.22)	60.0 (3.20)[a]	44.6 (3.36)[b]	26.9 (6.02)[b]
621	45.0 (3.20)[a]	51.7 (0.52)	44.2 (3.36)[b]	63.8 (8.90)[b]	50.2 (0.08)	61.7 (3.82)[b]	49.4 (0.20)	41.3 (4.56)[b]
622	49.8 (0.44)	52.7 (4.34)[b]	50.4 (0.22)	68.1 (16.2)[b]	49.6 (0.58)	61.0 (4.04)[b]	52.9 (2.26)	16.7 (16.6)[b]
623	36.9 (11.6)[b]	58.3 (11.3)[b]	41.9 (10.6)[b]	51.3 (2.14)	47.5 (2.14)	57.9 (10.2)[b]	46.3 (3.50)[b]	40.6 (11.4)[b]
624	42.3 (2.70)[a]	53.1 (1.42)	44.6 (3.74)[b]	54.6 (2.92)[a]	45.6 (1.66)	75.8 (7.56)[b]	– –	– –
AV.	38.9 (2.68)[a]	54.5 (1.63)	44.8 (2.01)	58.1 (2.52)[a]	47.3 (2.31)[a]	61.2 (4.35)[b]	48.1 (1.81)	30.9 (5.97)[b]

[a] Significantly different from 50 percent at the 5 percent significance level, 2-Tailed *t*-Test.

[b] Significantly different from 50 percent at the 1 percent significance level, 2-Tailed *t*-Test.

t-test). Six of the eight rats were under clear control of the contingencies as their choice frequencies decreased back toward their original levels, the two exceptions being rats 612 and 613. Thus, six of the eight subjects and the aggregate choice data show standard Allais-type violations of expected utility theory.

Tests of H2 in the southeast corner of the unit probability triangle show that it fails to hold. Rather, fanning in characterizes the data here. Under the initial implementation of condition 3, average choice of the more certain (E) alternative jumped to 58.1 percent ($t = 2.52$, $p < 0.05$, 2-tailed *t*-test). Equally as important, with the return to pair 1 choices, mean percentage choice of the certain alternative, A, fell to 47.3 percent ($t = -2.31$, $p < 0.05$, 2-tailed *t*-test). Choice was under reversible control, with a pattern of fanning in for six of the eight rats, 611, 613, and 621–624. The two remaining rats, 612 and 614, were unresponsive to the changes in experimental conditions, rather than displaying any hint of fanning out. A second implementation of condition 3 confirmed and strengthened these results, as subject 614 now showed clear fanning in, as well.[4]

The results of the sensitivity test are shown in the last column of Table 2. On average, the rats chose the more certain, dominated alternative (G) 30.9 percent of the time, well below the 50 percent mark. All the rats were responsive to the 1/8 chance of a 1 sec delay probability, as the mean choice of the dominated alternative was below 50 percent in all cases, and one can reject indifference between prospects for all subjects at the 1 percent significance level or better, including rat 612, who was largely unresponsive to the other experimental manipulations. These results show the rats to be sensitive to delay probabilities

[4] MacDonald, Kagel, and Battalio (1989) report comparable failures of fanning out for rats over gains (food rewards) in previously unexplored areas of the unit probability triangle (the northwest corner, in this case).

as small as 1/8. Hence, we cannot attribute the failure of H2 here to any inherent inability to recognize such low probability outcome events.

III. Experiment 2: Humans' Choices Over Losses

Studies of human choices over losses involved responses to a series of questions requiring subjects to indicate which of two gambles they preferred. All questions were of the following form:

Example 3

A: Losing $14 if 1–100
B: Losing $20 if 1–70
Losing $0 if 71–100
Answer: (1) I prefer *A*. (2) I prefer *B*. (3) Indifference

The numbers following each dollar amount referred to 100 numbered poker chips in a bowl sitting in front of the room, one of which would be drawn to determine payoffs. Thus, if a subject chose A, he or she would lose $14 with certainty as any numbered poker chip drawn gave this result. If, however, the subject chose B, and a chip numbered 1–70 was drawn, he or she would lose $20, but if a chip numbered 71–100 was drawn, the subject would lose $0.

Subjects were paid off on a *single* question, determined at random, through the draw of a poker chip, after all the questions were answered. They then drew a poker chip from the bowl of 100 chips to determine the actual payoff for that question. Immediately prior to answering these questions, subjects answered three hypothetical questions and went through the two-stage chip drawing procedure, to ensure their understanding of the payoff process. Subjects were given $30 cash balances against which to cover the possibility of losses. Payoffs were set so that all subjects would walk away with positive net earnings.[5]

Paying off subjects on a single question eliminates wealth effects from directly distorting preferences. However, papers by Charles Holt (1986) and Adi Karni and Zvi Safra (1987) show conditions under which answers to questions in a series of questions, only one of which will result in a payoff, will not be independent of the set of questions posed. This criticism applies to both the Gordon M. Becker, Morris H. DeGroot, and Jacob Marschak (1964) procedure for determining certainty equivalent values of gambles, and to simple elicitation of preferences over a series of paired alternatives, such as employed here. Responses to each question are independent of the set of questions posed if the independence axiom of expected utility theory holds, or subjects choose as if each gamble will actually be selected (as in the "isolation effect" in prospect theory). To test for independence, Colin Camerer (1989) has permitted subjects to change their choices once the question to be played out had been selected. Few subjects changed their choices. Answers reported here come from two separate series of questions, with several identical questions across the two series producing strong repeat reliability in the answers (Raymond C. Battalio, John H. Kagel, and Komain Jiranyakul, in press). This, too, is consistent with the independence assumption, as there were substantial differences in the other questions included in the two series. Since the data below show frequent violations of the independence axiom, we conclude that subjects treated each gamble separately, either in response to our urging them to do so or out of a natural isolation effect of the sort promoting the absence of asset integration reported in a number of choice experiments (Kahneman and Tversky, 1979; Camerer, 1989; Battalio, Kagel, and Jiranyakul, in press).[6]

[5] See Battalio, Kagel, and Jiranyakul (in press) for details of the experimental procedures employed.

[6] In the experiment using rats, subjects were paid off following each response. However, the payoffs, in terms of water reinforcement, were independent of the alternative chosen. In this case, there were no wealth effects in the sense that the outcomes of the prospects had no effect on the physical state of the rats.

Table 3—Choices Over Losses: Human Subjects

Set 1: Allais-Type Common Ratio Changes
$A_1(-\$14, 1.0)$ or $B_1(-\$20, 0.70)$
$A_2(-\$14, 0.20)$ or $B_2(-\$20, 0.14)$

Possible Choice Patterns	Patterns Consistent with	Choice Frequencies Set 1
A A	EU	10
B B	EU	5
B A	Fanning Out	10
A B	Fanning In	4

Set 2: New Tests of Fanning Out
2.1 $A_3(-\$20, 0.74; -\$14, 0.20)$ or $B_3(-\$20, 0.88)$
$A_4(-\$14, 0.90)$ or $B_4(-\$20, 0.63)$
2.2 $A_5(.-\$20, 0.60; -\$14, .40)$ or $B_5(-\$20, 0.88)$
$A_4(-\$14, 0.90)$ or $B_4(-\$20, 0.63)$

Possible Choice Patterns[a]	Patterns Consistent with	Choice Frequencies Set 2.1	Set 2.2
A A	EU	7	3
B B	EU	11	17
B A	Fanning out	1	5
A B	Fanning in	15	7

[a] Stochastically dominated alternatives listed first.

Table 3, set 1, shows one set of Allais-type common ratio alternatives employed (these are illustrated in Figure 2). The gambles employed are shown first, with the first number in brackets showing the dollar amount of the loss, followed by its probability (as a convention, we omit displaying zero loss probabilities). Below this are shown individual subject responses to the questions. The four possible choice patterns in moving from higher to lower probability losses are specified, along with their relationship to expected utility theory and the fanning out hypothesis. H2, as specified in Machina (1982, 1987), is a weak inequality that includes expected utility theory as a special case. Consequently, the fanning out hypothesis has power here to the extent that it organizes *deviations* from expected utility theory. It does this nicely, as a little under half the responses violate expected utility theory, and of these, 71.4 percent of the deviations correspond to fanning out rather than fanning in, far more than one would expect on the basis of chance factors alone.[7]

Choice sets 2.1 and 2.2. report tests of fanning out over losses in the southeast corner of the unit probability triangle (see Figure 2). In set 2.1, expected utility theory organizes 52.9 percent of the data compared to 50 percent on the basis of chance factors alone. Fanning out breaks down completely here, as fanning in organizes 44.1 percent of the data, well beyond expectations based on chance factors alone ($p < 0.01$).

Choice set 2.2 replaces prospect A_3 in set 2.1, with its 6 percent chance to escape without losses, with prospect A_5, which guarantees losses, should it be chosen. The certainty of losses reduces the attractiveness of the A alternative, thereby producing greater conformity with expected utility theory and increasing the frequency of clear fanning out, although fanning out still fails to explain a majority of the deviations from expected utility theory.

[7] Fanning out constitutes one of two possible deviations from expected utility theory. A binomial test statistic comparing observed frequencies with expected frequencies and a 10 percent significance level is employed in making these statistical evaluations.

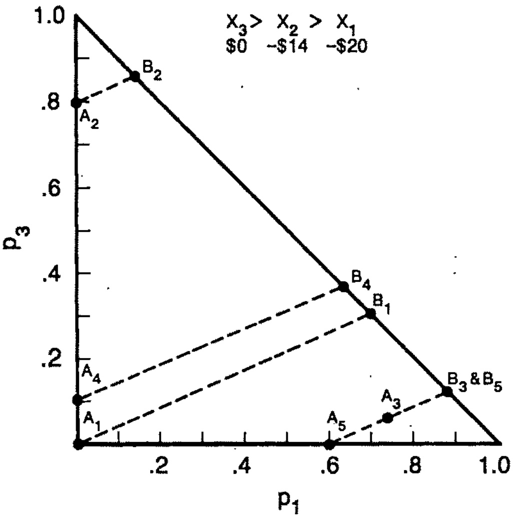


FIGURE 2. PROSPECTS EMPLOYED IN TESTS OF FANNING OUT WITH HUMAN SUBJECTS. PROSPECTS CORRESPOND TO THOSE SPECIFIED IN TABLE 3

The relatively superior performance of fanning out in the southeast portion of the unit probability triangle when the prospects are located on the edges of the triangle, compared to the interior, is characteristic of responses to other loss questions as well (Battalio, Kagel, and Jiranyakul, in press). It indicates that certain losses are much less attractive than uncertain losses with equal expected value when they provide some chance of escaping losses completely. Introducing a chance to escape losses on both alternatives promotes choice on the basis of the lowest loss alternative. This produces fanning out and classic Allais-type common ratio effects as in choice set 1, but it can also produce fanning in, as in choice set 2.1.[8]

IV. Summary and Conclusions

Results from two sets of experiments are reported, testing for fanning out of indifference curves over random prospects involving losses. For both rats and humans, classic Allais-type violations of expected utility theory are reported. However, new tests of fanning out in the southeast corner of the unit probability triangle show systematic fanning in over losses, contrary to Machina's (1982, 1987) H2. Lotteries in this corner of the loss triangle consist of highly positively skewed distributions over losses (i.e., most of the probability mass is on the left of the distribution, toward the maximum negative value), so that there is a "long left tail."

Other tests, over gains, for both human and animal subjects (Battalio, Kagel, and Jiranyakul, in press; MacDonald, Kagel, and Battalio, 1989) show comparable results: Allais-type violations of expected utility theory for standard manipulations, fanning in in previously unexplored areas of the unit probability triangle (in the northwest corner of the triangle in this case). These are largely negative results, as we do not offer an alternative model to explain the data. There are models that can do the job, however; in particular, prospect theory (Kahneman and Tversky, 1979) can organize these patterns of behavior (although it does not necessarily imply them). But more broad-based tests of these alternative theories show important deficiencies as well. (See Camerer, 1989; Battalio, Kagel, and Jiranyakul, in press; David W. Harless, 1987; and Chris Starmer and Robert Sugden, 1987, for comprehensive tests of a number of nonexpected utility models.) Hence, we hesitate to declare a winner in terms of which theory best organizes the data. Nevertheless, the results here are important, given the surprising power of H2 in organizing a number of established violations of expected utility theory (Machina, 1982, 1987) and the preference reversal phenomena (Safra, Segal, and Spivak, 1988). The results also establish some new stylized facts for theorists to organize in their efforts to develop alternatives to expected utility theory as a descriptive model of choice under uncertainty.

[8] Battalio, Kagel, and Jiranyakul (in press) report mirror image responses for tests of fanning out over gains in the northwest corner of the unit probability triangle as well: Prospects located on the edges of the unit probability triangle are much more likely to satisfy weak fanning out compared to cases where both alternatives involve some possibility of zero gains. John Conlisk (1989) reports similar results as well.

REFERENCES

Battalio, Raymond C., Kagel, John H. and MacDonald, Don N., "Animals' Choices Over Uncertain Outcomes: Some Initial Experimental Results," *American Economic Review*, September 1985, *75*, 597–613.

______, ______ **and Jiranyakul, Komain,** "Testing Between Alternative Models of Choice Under Uncertainty: Some Initial Results," *Journal of Risk and Uncertainty* (in press).

Becker, Gordon M., DeGroot, Morris H. and Marschak, Jacob, "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, July 1964, *9*, 226–32.

Camerer, Colin, "An Experimental Test of Several Generalized Utility Theories," *Journal of Risk and Uncertainty*, April 1989, *2*, 61–104.

Chew, Soo Hong and MacCrimmon, Kenneth R., "Alpha-Nu Choice Theory: A Generalization of Expected Utility Theory," University of British Columbia Faculty of Commerce and Business Administration Working Paper No. 669, mimeo., 1979.

Conlisk, John, "Three Variants on the Allais Example," *American Economic Review*, June 1989, *79*, 392–407.

Green, Leonard, "Self-Control Behavior in Animals," in V. L. Smith, ed., *Research in Experimental Economics*, Vol. 2, Greenwich, CT: JAI Press, 1982.

Grether, David M., and Plott, Charles R., "Economic Theory of Choice and the Preference Reversal Phenomena," *American Economic Review*, September 1979, *69*, 623–38.

Hagen, Ole, "Towards a Positive Theory of Preferences Under Risk,"in M. Allais and O. Hagen, eds., *Expected Utility Theory and the Allais Paradox*, Dordrecht, Holland: D. Reidel, 1979.

Harless, David W., "Predictions About Indifference Curves in the Unit Probability Triangle: A Test of Some Competing Decision Theories," working paper, Indiana University, 1987.

Holt, Charles, "Preference Reversals and the Independence Axiom," *American Economic Review*, June 1986, *76*, 508–15.

Kagel, John H. and Green, Leonard, "Intertemporal Choice Behavior: Evaluation of Economic and Psychological Models," in L. Green and J. H. Kagel, eds., *Advances in Behavioral Economics*, Vol. 1, Norwood, NJ: Ablex Publishing, 1987.

______, **MacDonald, Don. N., Green, Leonard and Battalio, Raymond C.,** "Risk Preferences Over Losses in Rats: Responses to Variable Shock Levels and Delays to Reinforcement," 1988, mimeo., University of Pittsburgh.

Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

Karni, Edi and Safra, Zvi, "'Preference Reversal' and the Observability of Preferences by Experimental Methods," *Econometrica*, May 1987, *55*, 675–85.

Killeen, Peter, "On the Measurement of Reinforcement Frequency in the Study of Preference," *Journal of the Experimental Analysis of Behavior*, May 1968, *11*, 263–69.

Lichenstein, Sarah and Slovic, Paul, "Reversals of Preference Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, July 1971, *89*, 46–55.

MacDonald, Don N., Kagel, John H. and Battalio, Raymond C., "Animals' Choices Over Uncertain Outcomes: Further Experimental Results," mimeo., Northeast Louisiana University, 1989.

Machina, Mark J., "'Expected Utility' Analysis Without the Independence Axiom, " *Econometrica*, March 1982, *50*, 277–324.

______, "The Economic Theory of Individual Behavior Toward Risk: Theory, Evidence and New Directions," IMSSS Technical Report no. 433, Stanford University, 1983.

______, "Choice Under Uncertainty: Problems Solved and Unsolved," *Journal of Economic Perspectives*, Summer 1987, *1*, 121–54.

MacCrimmon, Kenneth R. and Larsson, Stig, "Utility Theory: Axioms Versus 'Paradoxes', " in M. Allais and O. Hagen, eds., *Expected Utility Hypotheses and the Allais Paradox*. Dordrecht, Holland: D. Reidel, 1979.

Markowitz, Harry M., *Portfolio Selection: Efficient Diversification of Investments*, New Haven, CT: Yale University Press, 1959.

Mazur, James E., "Fixed and Variable Ratios and Delays: Further Tests of an Equivalence Rule," *Journal of Experimental Psychology: Animal Behavior Processes*, 1986, *12*, 116–24.

Mossin, Jan, "A Note on Uncertainty and Preferences in a Temporal Context," *American Economic Review*, March 1969, *59*, 172–74.

Safra, Zvi, Segal, Uzi and Spivak, Avia, "Preference Reversals and Nonexpected Utility Behavior," *CARESS* working paper no. 88-19, University of Pennsylvania, 1988.

Starmer, Chris, and Sugden, Robert, "Violations of the Independence Axiom: An Experimental Test of Some Competing Hypotheses," Economics Research Center discussion paper no. 24, University of East Anglia, 1987.

Preference Reversal and Nonexpected Utility Behavior

By ZVI SAFRA, UZI SEGAL, AND AVIA SPIVAK*

The preference reversal phenomenon, first discovered by Sarah Lichtenstein and Paul Slovic (1971), puzzled economists for a long time, as it seemed to indicate nontransitive preferences (see David Grether and Charles Plott, 1979). This phenomenon emerges when a decision maker prefers lottery A to lottery B, but sets a higher selling price on B than on A. Three recent works (Charles Holt, 1986; Edi Karni and Zvi Safra, 1987; and Uzi Segal, 1988) proved that this apparent nontransitive behavior strongly depends on the expected utility hypothesis and on the special mechanism used to find decision makers' selling prices of lotteries. When interpreted as a two-stage lottery with either the independence axiom or the reduction-of-compound-lotteries axiom violated, this mechanism does not elicit decision makers' true certainty equivalents of lotteries.[1,2]

Holt raised the question of a possible connection between the preference reversal phenomenon and other types of nonexpected utility behavior. This paper tries to answer that question. We show below that the same conditions that imply preference reversals also imply the Allais paradox (Maurice Allais, 1953) and the common-ratio effect (Kenneth MacCrimmon and Stig Larsson, 1979). In other words, we can explain the preference reversal phenomenon not only as a transitive nonexpected utility behavior, but also as consistent with other violations of this theory.

I. The Preference Reversal Phenomenon

Let $A=(x,p;y,1-p)$ be a lottery. This lottery yields x dollars with probability p and y dollars with probability $1-p$. We assume throughout that all prizes are bounded within a $[-M,M]$ segment. Let $\geq$ be the decision maker's preference relation over lotteries. The functional V represents this relation if $V(A)\geq V(B)\Leftrightarrow A\geq B$. The decision maker's certainty equivalent of A (CE(A)) is defined implicitly as that number z for which $(z,1)\sim A$. To find its value, Gordon Becker, Morris DeGroot, and Jacob Marschak (1964) suggested the following monetary incentive: Let the decision maker announce a selling price of A. Next, select at random an offer price from an $[a,b]$ segment that includes this announced price. The decision maker will win the offer price if it exceeds the selling price. On the other hand, if the offer price is less than the selling price, the decision maker will keep the lottery and play it.

Intuitively, decision makers' optimal strategy would be to announce their true certainty equivalents of A. Otherwise, they might be forced to sell the lottery while willing to keep it, for example, when CE(A) $=3$, the selling price is 1, and the offer price is 2. Alternatively, they might be forced to keep the lottery while willing to sell, for example, when CE(A) $=3$, the selling price is 5, and the offer price is 4. Nevertheless, carefully conducted experiments proved that subjects have a strong tendency to reverse their preferences, that is, to prefer a lottery A to B, but to set a higher selling price on B than on A (Lichtenstein and Slovic, 1971;

*Tel Aviv University, Tel Aviv, Israel; University of Toronto, Toronto, Canada; and Ben-Gurion University, Beer Sheva, Israel, respectively. This research was partially supported by the Social Sciences and Humanities Research Council of Canada. We are grateful to the referees for helpful comments.

[1] The main point of Holt's explanation is slightly different, since he is more concerned with the random lottery incentive mechanism that was used to control for wealth effects. In this paper we rather put the emphasize on the price selection mechanism. See Section I.

[2] Amos Tversky, Paul Slovic, and Daniel Kahneman (1990) have recently proposed a different explanation for the preference reversal phenomenon. They argue that what appear to be violations of transitivity may be a result of a bias introduced by the mode of reporting the prices.

Grether and Plott, 1979; Werner Pommerehne, Friedrich Schneider, and Peter Zweifel, 1982; and Robert Reilly, 1982). Such behavior violates transitivity, one of the most fundamental assumptions in economics.

To solve this paradox, Karni and Safra (1987) suggested a different interpretation of the Becker-DeGroot-Marschak mechanism. Rather than a simple lottery, they analyzed this mechanism as a two-stage lottery without the independence axiom. Let π be the decision maker's announced selling price of the lottery A. Lottery A is won with a probability of $(\pi - a)/(b-a)$, while a $(b-\pi)/(b-a)$ mass of probability is spread uniformly over the $[\pi, b]$ segment.

While accepting the reduction-of-compound-lotteries axiom, Karni and Safra proved that the decision maker's optimal announcement, $\pi(A)$, always equals the true certainty equivalent of A if and only if the subject is an expected utility maximizer. The preference reversal phenomenon thus becomes just more evidence against expected utility theory, but does not prove intransitivity. (For other transitive interpretations of this phenomenon, see Holt, 1986; and Segal, 1988).

The question that naturally follows is whether the preference reversal phenomenon is consistent with other evidence against expected utility theory.[3] In other words, does the preference reversal phenomenon belong to the large family of nonexpected utility behavioral patterns that can be uniformly explained within the standard nonexpected utility models (Mark Machina, 1982; John Quiggin, 1982; Chew Soo Hong, 1983; Eddie Dekel, 1986; and Menahem Yaari, 1987), or does it present us with an entirely different problem, not yet analyzed?

Our main conclusion here is that there does exist a connection between preference reversals and other nonexpected utility phenomena like the Allais paradox and the common-ratio effect. Moreover, almost the same condition as Machina's Hypothesis II, which he proved explained these paradoxes, also explains the preference reversal phenomenon. A similar result for a specific functional form (that of Quiggin and Yaari) that does not satisfy either Machina's Hypothesis II or our coming assumptions appears in Karni and Safra (1989).

To explain this hypothesis and its connection to the Allais paradox, let $x<y<z$ be fixed prizes and consider the set Δ of lotteries over these prizes, $\Delta = \{(x,p;y,1-p-q;z,q)\}$. Each such lottery is represented by a point in the unit triangle in Figure 1 (see Machina, 1987). It follows by monotonicity that, for a given value of p, a higher q benefits the decision maker because a probability mass shifts from the middle outcome to the high outcome. Similarly, for a given value of q, a lower p is preferred. It thus follows that the indifference curves in this triangle have a positive slope, with higher curves indicating a higher utility level. The expected utility of the lottery $(x,p;y,1-p-q;z,q)$ is $p[u(x)-u(y)]+q[u(z)-u(y)]+u(y)$; hence the slope of the indifference curves is $[u(y)-u(x)]/[u(z)-u(y)]$, and the indifference curves are parallel straight lines (Figure 1a).

Let $x=0$, $y=1000000$, and $z=5000000$, and consider the four lotteries $A=(x,0.9;z,0.1)$, $B=(x,0.89;y,0.11)$, $C=(x,0.01;y,0.89;z,0.1)$, and $D=(y,1)$ (Figure 1b). For A to be preferred to B and D to C, which is the Allais paradox, indifference curve β must be steeper than α. This is the "fanning out" effect, also known as Machina's Hypothesis II, which means that for $p'<p$ and $q'>q$, the slope of the indifference curve at (p',q') is higher than the slope at (p,q). (This refers, of course, to two different indifference curves.)

This hypothesis also explains the common-ratio effect, where for $x<y$, $(0,1-p;x,p)\sim(0,1-q;y,q)$, but for $\lambda<1$, $(0,1-\lambda q;y,\lambda q)>(0,1-\lambda p;x,\lambda p)$. For example, as Daniel Kahneman and Amos Tversky (1979) found (in a hypothetical experiment), $(3000,1)>(0,0.2;4000,0.8)$, but $(0,0.8;4000,0.2)>(0,0.75;3000,0.25)$. This occurs when the indifference curve through

[3] For a discussion of the connection between the preference reversal phenomenon and the common-ratio effect within a different approach, see Holt (1986).

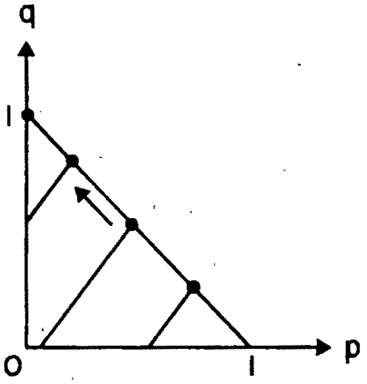


FIGURE 1a

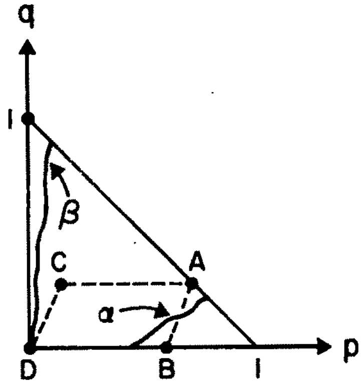


FIGURE 1b

(0,0) and above (0.2,0.8) is steeper than the one that goes through (0.75,0) and below (0.8,0.2).

This graphical representation holds only for triple-outcome lotteries, but the results hold for general lotteries as well. Let F and G be two cumulative distribution functions. We say that F dominates G by first-order stochastic dominance and denote it by $F \geq_1 G$ if, for every x, $F(x) \leq G(x)$. A preference relation satisfies the first-order stochastic dominance axiom if F is preferred to G whenever $F \geq_1 G$. For a general distribution function F, let U_F be the decision maker's local utility at F (see Machina 1982, 1987). For small changes in the distribution, the decision maker's behavior can be approximated by the expected utility of U_F. Formally,

$$V(G) - V(F) = \int U_F(x)\,dG - \int U_F(x)\,dF + o(\|G-F\|),$$

where, as before, V represents the preference relation $\succeq$.

Let $R^F(x) = -U_F''(x)/U_F'(x)$. Machina made the following assumption concerning this expression:

ASSUMPTION H2 (Hypothesis II): *If F dominates G by first-order stochastic dominance, then for every x, $R^F(x) \geq R^G(x)$.*

Note that in Figure 1(b) C dominates B by first-order stochastic dominance. We later show that H2 implies that the slope of the indifference curve at C is higher than the slope of the indifference curve at B. We will use a similar assumption to analyze the preference reversal phenomenon.

II. Necessary and Sufficient Conditions for Preference Reversals

A preference reversal is established whenever a decision maker prefers lottery A to B, but sets a higher selling price on B than on A. In most of the experiments that found such reversals, B was (almost) a mean-preserving spread of A. To relate preference reversals to mean-preserving spreads, we must hypothesize about the manner in which the local utility function changes when the decision maker moves from one distribution to a mean-preserving spread of that distribution. Let $F \geq_2 G$ if G is a mean-preserving spread of F and make the following assumption:

ASSUMPTION A2: *If G is a mean-preserving spread of F, then for all x, $R^F(x) \geq R^G(x)$.*

A2 is closely related to Machina's Assumption H2, where F was assumed to dominate G by first-order rather than second-order stochastic dominance. These two assumptions relate to changes in absolute risk aversion across distributions and should not be confused with the change in R^F induced by a change in x (Machina, 1982; Hypothesis I refers to such changes). Instead of Assumptions A2 and H2, one can make the following two assumptions:

ASSUMPTION A3: *If G is a mean-preserving spread of F, then for all x, $R^F(x) \leq R^G(x)$.*

ASSUMPTION H3: *If F dominates G by first-order stochastic dominance, then for all x, $R^F(x) \leq R^G(x)$.*

Our analysis requires a few more definitions. The function V represents *risk aversion* if $F \geq_2 G$ implies $V(F) \geq V(G)$. It represents *risk loving* if $F \geq_2 G$ implies $V(F) \leq V(G)$. It is *quasi-concave* (*convex*) if for $\alpha \in [0,1]$, $V(F) = V(G)$ implies $V(\alpha F + (1-\alpha)G) \geq V(F)$ $(V(F) \geq V(\alpha F + (1-\alpha)G))$. If it is both quasi-concave and quasi-convex, then it satisfies the betweenness axiom (Chew, 1983; Dekel, 1986). If V is quasi-concave (convex), then indifference curves in the triangle are convex (concave). Betweenness implies straight (but not necessarily parallel) indifference curves.

PROPOSITION 1: *Let the function V satisfy the betweenness axiom. If it represents risk aversion, then Assumption H_i holds if and only if Assumption A_i holds as well, $i = 2,3$. If V represents risk loving, then Assumption H_i holds if and only if Assumption A_j holds, $i \neq j$, $i,j \in \{2,3\}$.*

The various assumptions and Proposition 1 are easily interpreted in Machina's triangle. If $R^F > R^G$, then the slope of the indifference curve through F is greater than its counterpart through G (see Figure 2). Comparing the movement from F to F' with that from G to G', we see that, in the first, a greater probability of the high outcome must compensate for an increased probability of the low outcome. This is so because an increase in the probability of the highest outcome z contributes less to total utility, because of the higher risk aversion at F.

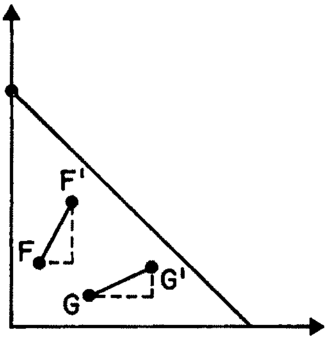


FIGURE 2

More formally,

$$(1)\quad \frac{\Delta q}{\Delta p}\bigg|_{V(G)=\text{const.}} = \frac{U_G(y)-U_G(x)}{U_G(z)-U_G(y)}.$$

This expression increases if G is replaced by F such that $R^F \geq R^G$, or equivalently, when $U_F(x) = \psi[U_G(x)]$, where ψ is an increasing concave function.

Proposition 1 is illustrated in Figure 3 for $H_2 \Leftrightarrow A_2$. Betweenness implies straight indifference curves. Unlike expected utility theory, they are not necessarily parallel. The expected value of the lottery $(x,p;y,1-p-q;z,q)$ is given by $\mu = p(x-y)+q(z-y)+y$. The locus of lotteries with a given mean μ is a straight line with a slope of $(y-x)/(z-y)$. It follows from (1) that, if the decision maker is risk averse (in which case the local utility function is concave), then for a general distribution F,

$$\frac{U_F(y)-U_F(x)}{U_F(z)-U_F(y)} > \frac{y-x}{z-y}.$$

Therefore, the slope of the indifference curves is higher than the slope of the line along which μ is a constant (see Figure 3). Because higher indifference curves represent higher utility levels, risk aversion is expressed by the fact that moving in the northeast direction on the line along which μ equals a constant, that is, affecting mean preserving spread, reduces the decision maker's utility.

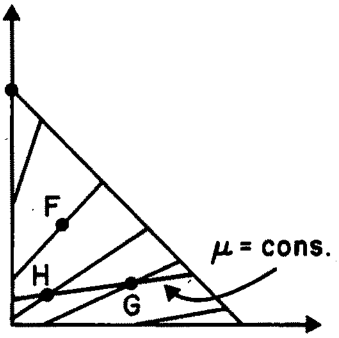


FIGURE 3

By Assumption H2, risk aversion (hence the slope of the indifference curves) at points above and to the left of G (for example, F) is higher than at G. By Assumption A2, risk aversion becomes higher as we move to the left on the line along which μ equals a constant (for example, to H). Obviously, if indifference curves are straight lines, then Assumption H2 holds if and only if Assumption A2 is satisfied. Using the same diagram for quasi-concave or quasi-convex preferences, we can show that, for lotteries with three outcomes, under risk aversion and quasi-concavity Assumption A2 implies Assumption H2. Also, under risk aversion and quasi-convexity H2 implies A2.

Having clarified the meaning of the assumptions, we can now state the major result of this paper. For each lottery A, let $\pi(A)$ be the decision maker's optimal value of π, the announced selling price of A. We assume henceforth that this maximum is unique (see the appendix). Let

$$L(A,\pi)=\frac{\pi-a}{b-a}A+\frac{b-\pi}{b-a}Q(\pi,b),$$

where $Q(\pi,b)$ is the uniform distribution over $[\pi,b]$. $L(A,\pi)$ is the (compound) lottery obtained from the lottery A by the Becker-DeGroot-Marschak mechanism when the decision maker declares the selling price π. If $A=(x,p;y,1-p)$, then $L(A,\pi)$ yields x with probability $p(\pi-a)/(b-a)$, y with probability $(1-p)(\pi-a)/(b-a)$, and a $(b-\pi)/(b-a)$ probability mass, uniformly spread over the $[\pi,b]$ segment. Let $U^A=U_{L(A,\pi(A))}$ and $U^B=U_{L(B,\pi(A))}$ be the local utilities of the lotteries obtained from A and B with A's optimal selling price.

THEOREM 2: *Let $A>_2B$.*

a. *If V represents risk aversion and satisfies Assumption A3, then preference reversals will not occur. A sufficient condition for a preference reversal is that $\int U^B\,dB>U^B(\pi(A))$.*

b. *If V represents risk loving and satisfies Assumption A2, then preference reversals will not occur. A sufficient condition for a preference reversal is that $\int U^B\,dB<U^B(\pi(A))$.*

The sufficient conditions indicate that preference reversals between A and B require a sizable change in the absolute measure of risk aversion between $L(A,\pi(A))$ and $L(B,\pi(A))$. To illustrate the inequality condition in Theorem 2(a), consider Figure 4. Here $A=(x,p;y,1-p)$. Note that, because in this example $U^A(x)=U^B(x)$ and $U^A(y)=U^B(y)$, $\int U^B\,dA=\int U^A\,dA$. The first-order condition for maximization is

$$\frac{\partial}{\partial\pi}V(L(A,\pi))\bigg|_{\pi=\pi(A)}=0.$$

As in the proof of Theorem 2 (see Appendix), this condition implies that $\int U^A dA=U^A(\pi(A))$. Let $\Delta=\int U^B d(A-B)$ $(\Delta>0)$. The condition now becomes $U^A(\pi(A))>U^B(\pi(A))+\Delta$. This implies that the local utility U^A is more concave than U^B, and the value it assigns to $\pi(A)$ is much larger than $U^B(\pi(A))$.

The conditions for existence or nonexistence of preference reversals somewhat resemble those leading to a Giffen good. Assume first that V represents risk aversion. The movement from A to B has two effects. Under Assumption A3 they work in the same direction, to reduce π, while under Assumption A2 they work in opposite directions. First, because B is worse than A, it will have a smaller π for the same local utility. Second, the local utility changes

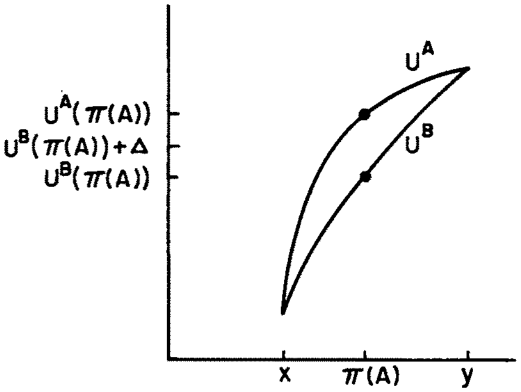


FIGURE 4

from $L(A,\pi(A))$ to $L(B,\pi(A))$. The sign of the change depends on whether Assumption A2 or A3 is satisfied. Under Assumption A3, since B is a mean-preserving spread of A, the consumer is more risk averse at $L(B,\pi(A))$ than at $L(A,\pi(A))$ and hence gives the risky prospect B an even lower certainty equivalent. Under Assumption A2, the certainty equivalent of B increases. Thus, under Assumption A3 risk averters never show preference reversals, while under Assumption A2 they might. A similar argument explains the case of risk loving.

Two opposing effects also appeared in all the experiments such as Allais', the common-ratio effect, and others. Machina (1982) proved that a sufficient deviation from expected utility, provided by the H2 assumption, implies these behavioral patterns. Strikingly, the same Assumption H2 also relates to the preference reversal phenomenon. Consider functionals satisfying betweenness. By Theorem 2, preference reversals do not occur under Assumption H3 but might occur if Assumption H2 is satisfied. A sufficient condition for its occurrence is the fast fanout of indifference curves.

This observation agrees with consistent experimental results displaying a fairly high percentage of preference reversals (50–60 percent) among those who preferred the safer option and a fairly low percentage of preference reversals (10–15 percent) among those who preferred the riskier one. Our explanation suggests that Assumption H2 is much more frequent among risk averters than among risk lovers. This conclusion is supported by experiments done by Marc McCord and Richard de Neufville (1984). In their experiments they tried to elicit utility functions using different approaches and found that their results contradicted the expected utility analysis. As shown by Machina (1987), their results indicate that Assumption H2 prevails for risk averters. A further look at their results indicates that, for risk lovers, Assumption H2 does not prevail. Our analysis thus agrees with these results as well.

We conclude this section by showing how for *any* given pair of lotteries A and B that satisfy $A \geq_2 B$, it is possible to build a functional displaying preference reversal between them. The functional represents risk aversion, satisfies Assumption A2, and can be smoothed to be Frechet differentiable.

Example. Let A and B be two given lotteries satisfying $A >_2 B$. Consider the expected value functional $\mu(F) = \int x dF(x)$. Clearly, for μ we have $\pi(A) = \pi(B) = \mu(A) = \mu(B)$. Define $m = \mu(L(A,\pi(A))) = \mu(L(B,\pi(B)))$ and note that $L(A,\pi(A)) >_2 L(B,\pi(B))$. Let u: $\mathbb{R} \to \mathbb{R}$ be a concave function such that N, defined by $N(F) = \int u(x)\,dF(x)$, satisfies (1) $N(A) > N(B)$ (this follows by risk aversion),

$$(2)\quad N\left(\tfrac{1}{2}L(A,\pi(B)) + \tfrac{1}{2}L(B,\pi(B))\right) = m,$$

and

$$(3)\qquad N(L(B,\pi)) < m \text{ for all } \pi.$$

Conditions (2) and (3) are possible since u can be affinely transformed. Finally, let

$$V(F) = \max\{\mu(F), N(F)\}.$$

Indifference curves of V are (informally) depicted in Figure 5, where the sets $\{L(B,\pi)\colon \pi \in \mathbb{R}\}$ and $\{L(A,\pi)\colon \pi \in \mathbb{R}\}$ also appear and Q is the uniform distribution over $[a,b]$.

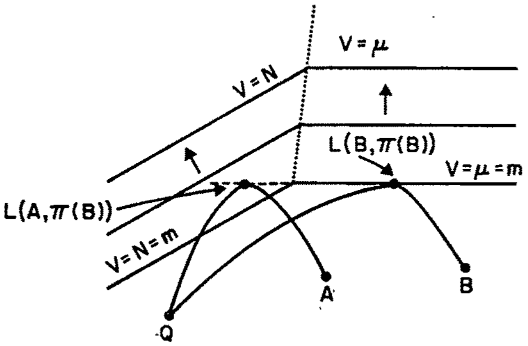


FIGURE 5

By construction we still have $\pi(B)=\mu(B)$; therefore the local utility function at $L(B,\pi(B))$ is linear. Furthermore, the local utility function at $L(A,\pi(B))$ is u and, by its concavity, $u(\pi(B))>\int u dA$. This condition is equivalent to the inequality condition in Theorem 2(a) (the roles of A and B were changed), which gives the existence of preference reversals between A and B ($V(A)>V(B)$ and $\pi(A)<\pi(B)$). Figure 5 clearly shows that the optimal point of $\{L(A,\pi)$: $\pi\in\mathbb{R}\}$ is to the left of $L(A,\pi(B))$, so $\pi(A)<\pi(B)$. Of course V does not represent strict risk aversion, but this can be achieved by replacing μ with an expected utility functional having a concave and almost linear utility function. Frechet differentiability of V results from smoothing out the kinks of V without affecting the existence of the preference reversal. Finally, V satisfies Assumption A2. Indeed, let $F\geq_2 G$ such that $V(G)=N(G)$ (this is the only interesting case). But then $N(F)\geq N(G)\geq\mu(G)$ (by risk aversion and the definition of V), thus $V(F)=N(F)$, and clearly $R^F=R^G$.

III. Summary and Conclusion

Preference reversals seem to violate transitivity, one of the most fundamental assumptions in economics. They occur when a decision maker prefers lottery A over B, but puts a higher selling price on B than on A. From the economist's point of view, these results are even more troublesome when subjects are offered a clear monetary incentive to reveal their true preferences. It was assumed that the Becker-DeGroot-Marschak mechanism creates this incentive. Karni and Safra (1987) and Segal (1988) showed that this is not the case if either the independence axiom or the reduction of compound lotteries axiom is violated. In this paper we have shown that, given Karni and Safra's interpretation, similar conditions to those implying the Allais paradox and the common-ratio effect also imply preference reversals. Holt (1986) also discussed such a connection, but his explanation is slightly different (see fn. 1). Furthermore, we have shown that this analysis also explains why the preference reversal phenomenon is more common among risk averters than among risk lovers.

This paper predicts that if risk-averse decision makers display preference reversals, they will also behave according to the Allais paradox and the common-ratio effect. The opposite, however, is not necessarily true. Further experiments can examine these predictions.

APPENDIX

PROOF OF PROPOSITION 1

We only prove that if V represents risk aversion, then H2 ⇔ A2. Assume that Assumption H2 is satisfied and take $F\geq_2 G$; thus, $V(F)\geq V(G)$. Let $\delta_x=(x,1)$ be the distribution that assigns all its mass to point x. $\delta_M=(M,1)$ dominates G by first-order stochastic dominance and is preferred to F; hence there exists $H\geq_1 G$ such that $V(H)=V(F)$. By betweenness, $R^H=R^F$ and by Assumption H2, $R^H\geq R^G$; hence $R^F\geq R^G$, which is Assumption A2.

Assume A2, let $F\geq_1 G$, and let $\mu(G)$ be the mean of G. If $V(\delta_{\mu(G)})\geq V(F)$, then let $G(\beta)=\beta G+(1-\beta)\delta_{\mu(G)}$ with $0\leq\beta\leq 1$ such that $V(G(\beta))=V(F)$. Since $G(\beta)\geq_2 G$, it follows by betweenness and Assumption A2 that $R^F=R^{G(\beta)}\geq R^G$.

Assume now that $V(\delta_{\mu(G)})<V(F)$. Without loss of generality, $G\neq\delta_{\mu(G)}$. (Otherwise, take $K\neq G$ such that $V(K)=V(G)$.) By risk aversion, $V(\delta_{\mu(G)})>V(G)$. Define $G_1=\max\{G+t_1,\delta_M\}$ such that $V(G_1)=V(\delta_{\mu(G)})$ ($F+t$ is a t-length rightward shift of F). Such t_1 exists because $M>\mu(G)$. Clearly, $V(\delta_{\mu(G_1)})>V(G_1)$. Continue to define $G_n=\max\{G_{n-1}+t_n,\delta_M\}$ such that $V(G_n)=V(\delta_{\mu(G_{n-1})})$ while $V(\delta_{\mu(G_n)})\geq V(G_n)$. The sequence $\mu(G_n)$ is increasing and converges to δ_M. (Otherwise, $G_n\rightarrow K$ and $V(\delta_{\mu(K)})=V(K)$ while $K\neq\delta_{\mu(K)}$, a contradiction.) Thus, for a sufficiently large n, $V(\delta_{\mu(G_n)})\geq V(F)\geq V(G_n)$ and, by Assumption A2, $R^{G_n}=R^{\delta_{\mu(G_{n-1})}}\geq R^{G_{n-1}}\geq\cdots\geq R^G$. Define $G_n(\beta)=\beta G_n+(1-\beta)\delta_{\mu(G_n)}$ and continue as before.

Uniqueness of the Local Maximum in the Case of Quasi-Concavity. We show that if V is quasi-concave, then $d/d\pi\, V(L(A,\pi))|_{\pi=\pi^*}=0$ implies that π^* is the global maximum; that is, $\pi^*=\pi(A)$.

Consider the expected utility functional given by $\mathrm{EU}(F)=\int U_{L(A,\pi^*)}\,dF$. By first-order conditions, $\mathrm{EU}(A)=\mathrm{EU}(\delta_{\pi^*})$; thus $\pi^*=\mathrm{CE}(A)$ for the local utility function $U_{L(A,\pi^*)}$. π^* is a global maximum of this function; hence $\int U_{L(A,\pi^*)}\,dL(A,\pi^*)>\int U_{L(A,\pi^*)}\,dL(A,\pi)$ for all π. If π^* is not a maximum of V, then there is $\pi\neq\pi^*$ such that $V(L(A,\pi))>V(L(A,\pi^*))$. But then, by quasi-concavity, $\int U_{L(A,\pi^*)}\,d[L(A,\pi)-L(A,\pi^*)]>0$, which is a contradiction.

PROOF OF THEOREM 2

a. *Necessity*. Let V satisfy Assumption A3. Because of risk aversion, $V(A)>V(B)$ and preference reversals do not occur if $\pi(A)>\pi(B)$. Following Karni and Safra (1987), let $\partial L(B,\pi)$ be the derivative of the path $L(B,\pi)$ with respect to its second argument, evaluated at π. By definition

$$\partial L(A,\pi)(x)=\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}[L(A,\pi+\varepsilon)-L(A,\pi)](x)$$

$$=\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\left[\frac{\pi+\varepsilon-a}{b+a}-\frac{\pi-a}{b-a}\right]A(x)$$

$$+\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\left[\frac{b-(\pi+\varepsilon)}{b-a}E(\pi+\varepsilon,b)\right.$$

$$\left.-\frac{b-\pi}{b-a}E(\pi,b)\right](x).$$

For $x<\pi$, the last element is zero, and the value is $A(x)/(b-a)$. For $x>\pi$, the second element is $-1/(b-a)$ (since $E(\pi+\varepsilon,b)(x)=(x-(\pi+\varepsilon))/(b-(\pi+\varepsilon))$, and the value is $(A(x)-1)/(b-a)$. It thus follows that $\partial L(A,\pi)=(A-\delta_\pi)/(b-a)$; hence

$$\left.\frac{\partial}{\partial\pi}V(L(A,\pi))\right|_{\pi=\pi(A)}$$

$$=\int U_{L(A,\pi(A))}\,d[\partial L(A,\pi(A))]$$

$$=\frac{1}{b-a}\int U_{L(A,\pi(A))}\,d[A-\delta_{\pi(A)}].$$

First-order conditions for maximum imply that this last expression equals zero. By Assumption A3 and because by the definition of L, $L(A,\pi(A))>_2 L(B,\pi(A))$, it follows that $\int U_{L(B,\pi(A))}\,d[A-\delta_{\pi(A)}]<0$. We thus obtain

$$\left.\frac{d}{d\pi}V(L(B,\pi))\right|_{\pi=\pi(A)}$$

$$=\frac{1}{b-a}\int U_{L(B,\pi(A))}\,d[B-\delta_{\pi(A)}]$$

$$=\frac{1}{b-a}\left[\int U_{L(B,\pi(A))}\,d[B-A]\right.$$

$$\left.+\int U_{L(B,\pi(A))}\,d[A-\delta_{\pi(A)}]\right].$$

By risk aversion, the first term is negative, and by Assumption A3, the second one is negative as well. Hence $d/d\pi\, V(L(B,\pi))|_{\pi=\pi(A)}<0$, which implies $\pi(A)>\pi(B)$.

Sufficiency. As we proved above, the sufficient condition is equivalent to

$$\left.\frac{\partial}{\partial\pi}V(L(B,\pi))\right|_{\pi=\pi(A)}>0;$$

hence it implies that $\pi(B)>\pi(A)$.

b. Let V satisfy Assumption A2. By risk loving, $V(B)>V(A)$ and $\int U_{L(B,\pi(A))}\,d[A-\delta_{\pi(A)}]>0$. This implies that $d/d\pi\, V(L(B,\pi))|_{\pi=\pi(A)}>0$; thus $\pi(B)>\pi(A)$. The sufficiency is proved as in part (a). □

REFERENCES

Allais, Maurice, "Le Comportement de l'Homme Rationnel Devant le Risque: Critique des Postulates et Axiomes de l'Ecole Americaine," *Econometrica*, October 1953, *21*, 503–46.

Becker, Gordon M., DeGroot, Morris H. and Marschak, Jacob, "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, July 1964, *9*, 226–32.

Chew Soo Hong, "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox," *Econometrica*, July 1983, *51*, 1065–92.

Dekel, Eddie, "An Axiomatic Characterization of Preferences Under Uncertainty: Weakening the Independence Axiom," *Journal of Economic Theory*, December 1986, *40*, 304–18.

Grether, David M. and Plott, Charles R., "Economic Theory of Choice and the Preference Reversal Phenomenon," *American*

Economic Review, September 1979, *69*, 623–38.

Holt, Charles A., "Preference Reversals and the Independence Axiom," *American Economic Review*, June 1986, *76*, 508–15.

Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

Karni, Edi and Safra, Zvi, "'Preference Reversal' and the Observability of Preferences by Experimental Methods," *Econometrica*, July 1987, *55*, 675–85.

______ **and** ______, "Rank-Dependent Probabilities," *Economic Journal*, forthcoming, 1990.

Lichtenstein, Sarah and Slovic, Paul, "Reversal of Preferences Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, July 1971, *89*, 46–55.

MacCrimmon, Kenneth R. and Larsson, Stig, "Utility Theory: Axioms Versus 'Paradoxes,'" in Maurice Allais and Ole Hagen, eds., *Expected Utility Hypotheses and the Allais Paradox*, Dordrecht, Holland: D. Reidel, 1979.

Machina, Mark J., "'Expected Utility' Analysis Without the Independence Axiom," *Econometrica*, March 1982, *50*, 277–323.

______, "Choice Under Uncertainty: Problems Solved and Unsolved," *Journal of Economic Perspectives*, Summer 1987, *1*, 121–54.

McCord, Marc and de Neufville, Richard, "Utility Dependence on Probability: An Empirical Demonstration," *Large Scale Systems*, February 1984, *6*, 91–103.

Pommerehne, Werner W., Schneider, Friedrich and Zweifel, Peter, "Economic Theory of Choice and the Preference Reversal Phenomenon: A Reexamination," *American Economic Review*, June 1982, *72*, 569–74.

Quiggin, John, "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization*, December 1982, *3*, 323–43.

Reilly, Robert J, "Preference Reversal: Further Evidence and Some Suggested Modifications in Experimental Design," *American Economic Review*, June 1982, *72*, 576–84.

Segal, Uzi, "Does the Preference Reversal Phenomenon Necessarily Contradict the Independence Axiom?" *American Economic Review*, March 1988, *78*, 233–36.

Tversky, Amos, Slovic, Paul and Kahneman, Daniel, "The Causes of Preference Reversal," *American Economic Review*, March 1990, *4*, 204–17.

Yaari, Menahem E., "The Dual Theory of Choice Under Risk," *Econometrica*, January 1987, *55*, 95–115.

Economic and Political Foundations of Tax Structure: Comment

By HERBERT J. KIESLING*

The paper by Walter Hettich and Stanley Winer (1988) on the economic and political structure of tax analysis is an important contribution, introducing a refreshingly positive approach into an analytical tradition where normative judgments have long been overly important. But Hettich and Winer pursue the logic of their approach only part of the way. In this comment, I would like to extend the Hettich-Winer (HW) positive approach to what I consider to be its logical conclusion, which should allow for a more complete analysis that avoids some possible methodological and ethical biases.

The HW model features voter support of a government where support is based positively upon the services received from a pure public good and negatively by tax costs including deadweight loss. Tax costs are affected by the choice of tax base, and different tax bases may involve different behavior responses (to avoid taxes) by different voters. The government wishes to maximize expected support, written as

$$(1) \qquad [b_i(G) - c_i(V_i)],$$

where G is the level of expenditures on all public goods, b_i is the increase in probability of a yes vote by voter i because of G, v_i is the cost in terms both of the tax payment t_i and the deadweight loss $d_i(v_i = t_i + d_i)$, and $(-c_i)$ is the decrease in the probability of a yes vote because of the cost. The amount of tax revenue the government can raise is in part dependent upon how much taxpayers change their behavior to avoid taxation. Such behavior is affected by the tax rate plus those taxpayer characteristics that affect how much taxes will affect the amount of tax-base activity he or she will supply. Given the existence of numerous possible tax-base activities, when the government maximizes its expected support subject to available tax revenue and the behavioral characteristics of voters in supplying tax-base activity, it is found that "the politically optimal tax structure requires marginal political opposition per dollar of tax revenue to be equalized across taxable activities for each taxpayer, as well as to be equalized across taxpayers for each activity."[1] In the absence of administrative costs, the optimal tax structure would be to tailor taxes specifically to the characteristics of each individual voting unit. With administrative costs this is not possible, but the key idea remains that optimally the tax system should accommodate to individual voter circumstances when the benefits of doing so outweigh the administrative costs.

The extension I suggest involves introduction of public goods more directly. With their G variable, HW only allow their model to consider the total expenditure level on public goods and not its composition, a procedure that can lead to analytical and ethical distortion.[2] In being concerned about the level of support for its tax system, the government will need to consider more detailed features of public expenditures so long as voter tax-base activity responses might be related to the public good mix that is provided, as is likely. Not only are there relations of complementarity and substitutability between public goods and different tax base activities, but voters will have

*Professor of Economics and Public/Environmental Affairs, Indiana University, Bloomington, IN 47405. I wish to thank Dave Wildasin and the referees for useful comments.

[1] Hettich and Winer, 1988, p. 705.

[2] To be fair, HW briefly consider the possibility of bringing in the benefit side in their fn. 16. On the other hand, in presenting their model, they present justifications for separating taxes from expenditures (p. 703).

differing amounts of "tax morale" according to how they perceive public good benefits. Economists would never consider abandoning the marginal utility to price ratio as a key criterion for private good purchases, and there is no reason why in principle the public sector should not be seen as working in the same way.

When the variability of government service mix and amount is added to the HW model, we have the government maximizing a support function,

$$\sum_{i}^{N}\sum_{k}[(b_i(G_k))-c_i(V_i)], \tag{2}$$

where the G_k are public goods and the other variables are as above.

The government faces the constraint

$$\sum G_k-\sum t_iB_i=0, \tag{3}$$

where the t_i are tax rates and B_i is the amount of taxable activity of the voter (tax base). With addition of multiple public goods into the model, and following HW otherwise, the B_i are influenced as

$$B_i=B_i(t_i,X_i,G_k). \tag{4}$$

Voters adjust their tax bases because of the tax rates, their personal characteristics concerning tax-base supply behavior (the x_i), and the amount and kind of public good they perceive their taxes to be buying.

Maximizing (3) subject to (4) with respect to the G_k and t_i gives

$$\frac{\partial B_i}{\partial G_k}=\lambda \text{ all } k \tag{5}$$

$$=\frac{\dfrac{\partial C_i}{\partial V_i}\dfrac{\partial V_i}{\partial T_i}}{B_i+t_i\dfrac{\partial B_i}{\partial T_i}+T_i\sum_k\dfrac{\partial B_i}{\partial G_k}}.$$

The politically optimal tax structure has tax rates that equalize marginal political costs per dollar of additional revenue across taxpayers, where this includes effects of the popularity or unpopularity of the public goods being provided.

Finally, the above does not take into account the fact that each voter has several tax bases (indexed by j) against which taxes are levied. Accounting for this gives

$$\lambda=\frac{\dfrac{\partial C_i}{\partial V_i}\dfrac{\partial V_i}{\partial t_{ij}}}{\left\{B_{ij}+t_{ij}\dfrac{\partial B_i}{\partial T_{ij}}+t_{ij}\sum_k\dfrac{\partial B_{ij}}{\partial G_k}\right\}\left\{\sum_{h\neq j}t_{ih}\dfrac{\partial B_{ih}}{\partial t_{ij}}+\sum_k t_{ih}\sum_k\dfrac{\partial B_{ih}}{\partial G_k}\right\}}. \tag{6}$$

Following HW, the second term in the denominator of the RHS of (6) is meant to capture interaction effects between taxes levied on one tax base and behavior adjustments in other tax bases. The second part of the expression allows for the possibility that these interactions might be affected by the nature of public expenditures.[3]

If the HW approach is correct, politicians pay attention to these tax bases to benefit interactions as they seek to win elections. Such interactions fall into at least two categories. First, there is a large variety of technical interactions between public services

[3] An interesting alternative way to illustrate the importance of this relationship is shown by Wildasin (1984) in discussing the implications of the following relationship between taxed goods or factors (tax bases) and government services:

$$\frac{\partial B_\iota}{\partial G_k}=\frac{\partial B_i^c}{\partial G_k}+MRS_k\frac{\partial B_i}{\partial I},$$

where the first term on the RHS is the income-compensated effect of G_k on the tax base and the second term is the marginal benefit of public good k times the income derivative of the supply function for B_i. But it is not possible for both $\partial B_i^c/\partial G_i$ and $\partial B_i/\partial B_k$ to equal zero if MRS_k is significantly positive (as it must be if the public good is not a waste) and if $\partial B_i/\partial I$ is very much different from zero (as it is for labor supply, for example). In a world full of existing taxes, all of them to some extent distortionary, changes in government expenditure cannot help but change tax base behavior in ways that affect welfare gains and losses, effects that may be far from negligible (welfare changes plausibly in excess of $0.80 per dollar of expenditure), Wildasin, 1984, p. 240.

and tax-base activities. These are well described in recent papers by David Wildasin (1984) and Assar Lindbeck (1982). Wildasin shows how a theoretical welfare maximization for public goods includes interactions of tax bases with public goods and presents simulations that show that for some activities, labor in particular, the public service to tax-base interaction can be quantitatively important.[4] Lindbeck presents a detailed exploration of the interaction of four types of public expenditures (transfers, subsidies to goods, goods with externalities, pure public goods) with labor supply, where he traces income and substitution effects from both the taxing activities and the publically provided goods. He argues convincingly that many of these activities have been of considerable historical importance.[5]

The second reason for considering tax-expenditure interactions involves the fact that tax-base activity depends importantly upon perceived benefits or public goods and services. While it is true that citizens with the sense of public spiritedness possessed by angels would gladly cooperate fully in their tax-base and compliance activities even when paying for goods and services they voted against, this is too much to expect from ordinary humans. There was a highly respected tradition among tax experts in the nineteenth and early twentieth centuries concerning the value of using "differentiation" of tax bases for this reason, where "objects of taxation are judiciously diversified in such a manner as to realize the ends desired," as C. F. Bastable (1895) put it.[6] If HW are correct, political decision makers sensitive to voter reactions will look carefully at perceptions of benefits on the part of the public as they design tax policies. This need not require that benefits be measured in some objective fashion but only that decision makers perceive benefit patterns and are able to impress the public with the reasonableness of their perceptions.

The accuracy of the HW view of the world, as amended here, is an empirical matter, of course, although the idea that voters take both sides of the fisc into account in considering the records of their elected representatives in tax matters seems a reasonable enough proposition. Nor is it an approach that has gone unnoticed in contemporary policy discussions. Thus the director of the Congressional Budget Office, Robert Reischauer (1988), in a recent paper widely cited in the press, has proposed as a way for the federal government to deal with current budget problems that taxes be installed that are earmarked for specific types of federal activities known to command widespread public support, doing this by shaping the taxes to fall on those getting benefits. Proposals specifically suggested by Reischauer include a broad-based energy tax or combination of gasoline taxes and oil import fees to support environmental programs such as acid-rain relief, a new value added tax to support extended Medicare and Medicaid programs, and the taxing of Social Security income to support the Supplemental Security Income program.[7]

[4] Wildasin shows that the proper Samuelsonian solution to optimal public good levels, instead of the traditional $MRS = MRT$ equality, is

$$MRS = \frac{MRT - \Sigma t_i \frac{\partial s_i}{\partial z}}{1 + \Sigma \frac{t_i}{q_i} \varepsilon_i},$$

where z is the publicly provided good, the x_i are tax-base activities and t_i; q_i, and ε_i are tax rates, gross of tax prices, and demand elasticities, respectively. For purposes of the HW model, these relationships are important if taken into account by the voters.

[5] For example, a publicly provided or subsidized good can be either complementary to leisure or a substitute for leisure according to the degree to which the leisure time is being used for home production of a competing good to the publicly provided good. For another example, if a pure public good is considered beneficial, a family might view its provision as raising its income, with associated income effects upon labor supply, lowering it if the worker is on the backward-bending portion of his supply curve.

[6] Bastable, 1895, p. 334.

[7] A drawback of using this amended HW approach is the complexity that it adds. While important, lack of complexity is only one goodness criterion; there are many others. Many proposals such as those of Reischauer (1988) do not add undue complexity to the tax code.

More generally, policies suggested by the amended HW approach would include actions where the Congress would adjust tax changes to current emphases in expenditure changes: In periods such as the Reagan years when the defense establishment was getting most of the emphasis, these costs could be financed relatively more by taxes falling upon those citizens who are "hawks"; when the emphasis changes to increases in services of compassion, taxes could be designed to fall relatively more on those who consider such social goals important. Welfare recipients, who cannot be taxed directly, might be easily persuaded to pay taxes in the form of in-kind payments of labor time. Public opinion polls, if carefully constructed, would be quite useful in facilitating this process by helping politicians learn voter preferences, and thus potential votes, with more precision.[8]

[8] If the amended HW model turns out to be a reasonably correct view of the world, then it follows that the analytical work of the "expert" students of the taxation process should reflect this, as was suggested in a paper by Henry Aaron a number of years ago (1969). Among other things, this would require that tax economists abandon their long exposition of the income tax against a comprehensive tax base as the "optimal" tax policy. For recent statements of this supposedly 'orthodox' position, see Charles McClure and G. Zodrow (1987) and Joseph Pechman (1987). Also see Richard Musgrave (1967).

REFERENCES

Aaron, Henry, "What Is a Comprehensive Tax Base Anyway?" *National Tax Journal*, December 1969, *22*, 543–49.

Bastable, C. F., *Public Finance*, London: Macmillan, 1895.

Browning, Edgar, "The Marginal Cost of Public Funds," *Journal of Political Economy*, April 1976, *84*, 283–98.

Hettich, Walter and Winer, Stanley, "Economic and Political Foundations of Tax Structure," *American Economic Review*, September 1988, *78*, 701–12.

Lindbeck, Assan, "Tax Effects Versus Budget Effects on Labor Supply," *Economic Inquiry*, October 1982, *20*, 473–89.

McClure, Charles and Zodrow, G., "Treasury I and the Tax Reform Act of 1986: The Economics and Politics of Tax Reform," *Journal of Economic Perspectives*, Summer 1987, *1*, 37–58.

Mill, John Stuart, *The Principles of Political Economy*, W. Ashly, ed., New York: Longmans-Green, 1926.

Musgrave, Richard, "In Defense of an Income Concept," *Harvard Law Review*, November 1967, *81*, 44–62.

Pechman, Joseph, "Tax Reform: Theory and Practice," *Journal of Economic Perspectives*, Summer 1987, *1*, 11–28.

Reischauer, Robert, "Thinking Systematically About the Budget Deficit Decisions Facing the Next President," Washington: Brookings Institution, 1988.

Rosenbaum, David, "At Helm, A Centrist Respected by His Foes," *New York Times*, March 13, 1989, p. 10.

Samuelson, Paul, "Pure Theory of Public Finance and Taxation," in J. Margolis and H. Guitton, eds., *Public Economics*, New York: St. Martins, 1969.

Wildasin, David, "On Public Good Provision with Distortionary Taxation," *Economic Inquiry*, April 1984, *22*, 227–43.

Technological Stagnation, Tenurial Laws, and Adverse Selection: Comment

By NADEEM NAQVI*

In a recent article in this *Review*, Kaushik Basu (1989) has made an important contribution to the literature on technological stagnation resulting from underinvestment or reduced adoption of new technology as an outcome of certain tenurial laws in many less developed countries. Tenurial laws in several less developed countries mandate a tenant's right to continue occupancy of a property (land or apartment) indefinitely except under a very limited set of circumstances;[1] the tenant, on the other hand, is prohibited by the same law from selling his tenancy rights. Under the law, the tenancy rights with fixed rental agreement are prior to, and nullify, any subsequent contract between the tenant and the landlord should the tenant decide to renege. In effect, the land tenure laws in these countries prohibit landlords from writing quit-contingent tenure contracts.[2]

As possible explanations of the typical lack of innovation in these countries, arguments have been put forward as to why it is not in the interest of the landlord (Gale Johnson, 1950) or why it is not profitable for the tenant (Michal Kalecki, 1976) to adopt an innovation.[3] These arguments only demonstrate why it is not in the interest of one of the parties to innovate, and, it can be shown that they inadvertently end up providing an explanation of why the other party would, in fact, adopt the innovation. Such arguments are clearly inadequate for explaining stagnation. Kaushik Basu (1989), on the other hand, has provided an explanation as to why both the landlord and the tenant would not find it worth their while to adopt the innovation under such tenurial laws. While this is an important step toward providing a complete theory of technological stagnation, Basu's argument is adverse-selection based, and it applies only to a landlord who owns *several* tenant-occupied rental properties. It does not explain, however, why a landlord with a *single* tenant-occupied rental property and his tenant do not innovate (p. 252). Without detracting from the importance of Basu's explanation, which I believe is compelling in the context of a multiple property holdings case, it may be pointed out that the inability of his theory to explain stagnation among single-property owners and their tenants is a significant shortcoming, for a large proportion of the landlords in many less developed countries own only a single rental property, and these properties again are characterized by a similar lack of innovation. It remains to be explained, therefore, why both a landlord with a single tenant-occupied rental property and the tenant do not find it worth their while to innovate, despite the innovation being such as to add more to property value than its cost. The purpose of this paper is to do just this. I provide a simple alternative and, I believe, a compelling explanation for the stylized fact.

I demonstrate how an attitude of aversion to risk on the part of the landlord, or the tenant, or possibly both can eliminate all innovation-cost-sharing contracts that the two parties could otherwise have entered into. The risk-aversion-based explanation of technological stagnation applies equally well

*Department of Economics, University of Georgia, Athens, GA 30602. I am grateful to a referee for encouragement and valuable suggestions. Any remaining errors are my sole responsibility.

[1] In India, for instance, if a landlord owns only one house and if he can "prove" that he requires it for his own use, then he can usually have an eviction order served to his tenant.

[2] I am indebted to the referee for bringing my attention to this feature of the land tenure laws.

[3] Also see Pranab Bardhan (1984), Amit Bhaduri (1973), Avishay Braverman and Joseph Stiglitz (1986), and David Newbery (1975), among others.

to both single- and multiple-property owners and their respective tenants. This alternative explanation also points to the fact that there are several reinforcing reasons for the typical lack of innovation in less developed countries; included in this list of reasons are the adverse selection problem and the problem arising from the agents' attitude of aversion to risk, which together with the property rights laws in these countries, contrive to produce a suboptimal outcome in terms of investment and innovative activity, and hence technological stagnation.

I. Risk Aversion and Technological Stagnation

Following Basu, I shall take an innovation to be an ordered pair (X,C), where X is the output consequent upon the innovation, and C is its cost. Considering only those innovations that add more to the value of the property than their cost, I impose the restriction that $X-C>0$; that is, I consider only those innovations that are socially profitable.[4]

Let q be the exogenously given probability that the tenant quits after the adoption of the innovation. This probability is usually greater than zero.[5] If the tenant quits, the landlord appropriates the entire output X, or the additional rent of X from a new tenant. If the tenant continues to occupy the property, he appropriates the entire output X. The landlord is not permitted to write quit-contingent tenure contracts. Such is the law.[6]

Assume that the landlord (L) offers a contract to the tenant (T) such that the landlord pays a fraction l of the cost of the innovation, with the remainder $(1-l)C$ to be paid by the tenant.[7] Also assume for the moment that both the landlord and the tenant are risk neutral. Then the value of the innovation to the landlord is

$$V_L(l,q)=qX-lC, \tag{1}$$

and the value of the innovation to the tenant is

$$V_T(l,q)=(1-q)X-(1-l)C, \tag{2}$$

where V_L and V_T are equal to the expected profits of the landlord and the tenant, respectively.

Let $X/C=1+r$, so that $r=(X-C)/C>0$ under the assumption that the innovation is socially profitable. For a given quit probability q, the landlord will find all cost-sharing contracts acceptable as long as $V_L(l,q)\geq 0$, and the set of all such contracts is

$$\bar{L}=\{l\colon l\leq(1+r)q,\quad r>0,\quad 0\leq q\leq 1\}. \tag{3}$$

The tenant, on the other hand, will accept all cost sharing contracts such that $V_T(l,q)\geq 0$, and these are

$$\bar{T}=\{l\colon l\geq -r+(1+r)q,\quad r>0,\quad 0\leq q\leq 1\}. \tag{4}$$

For a quit probability q, the equality in (3) determines the maximum cost share that the landlord is willing to bear, $l_L=(1+r)q$. Similarly, the equality in (4) determines the

[4] Basu (1989) also considers only those innovations for which $X-C>0$.

[5] Tenant quit probability is greater than zero for various reasons. Job change or transfer is one reason. Another common reason in the case of India is tenant harassment by the landlord.

[6] This is certainly a facet of the Delhi Rent Control Act of 1958, as noted by Basu. The tenancy laws in other cities in India are roughly the same. Further, instead of the landlord's output distribution $(X,0;q,1-q)$, one could have $(X_1,X_2;q,1-q)$, with $X_i>0$, and the argument developed here will go through as long as, given an output distribution, no further output sharing between the landlord and the tenant is possible under the law.

[7] If whoever undertakes the innovation has to bear its full cost, and if both the landlord (L) and the tenant (T) are risk neutral, then neither may innovate. For $V_L=qX-C$, the value of the innovation to the landlord, which is his expected profit, may be negative even if $X-C>0$. Similarly, $V_T=(1-q)X-C$ may be negative even if $X-C>0$. On this see Basu (1989, p. 252).

minimum landlord's cost share that the tenant is willing to accept, $l_T = -r + (1+r)q$. Further, since

$$(5) \qquad l_L - l_T = r = \frac{X-C}{C} > 0,$$

because the innovation is socially profitable, it follows that the maximum cost share that the landlord is willing to bear is greater than the minimum landlord's cost share acceptable to the tenant. Therefore the set of all cost-share contracts acceptable to all parties

$$(6) \qquad \Lambda = \bar{L} \cap \bar{T} = [l_T, l_L]$$

is unambiguously nonempty.[8] Hence the innovation will definitely be undertaken. In fact, the innovative activity in this economy will be socially optimal, despite the tenurial law that seemingly discourages innovation.

I show next that if at least one of the agents—the landlord or the tenant—is sufficiently risk averse, then the set Λ of cost-sharing contracts is empty. In order to develop this argument, it is instructive to present the above argument for the risk-neutral agents graphically. In Figure 1, OC (or line L) represents for each quit probability q the landlord's maximum acceptable cost share l_L; all cost shares *below* OC are acceptable to the landlord. For instance, for $q = q^0$, the landlord is willing to accept all costs shares $l \le l_L^0$. The segment AB (or line T), on the other hand, represents for each q the minimum landlord's cost share l_T acceptable to the tenant; all cost shares *above* AB are acceptable to the tenant. In particular, for $q = q^0$, since the tenant finds all $l \ge l_T^0$ acceptable, the set of mutually acceptable contracts is $\Lambda(q^0) = DE$. And this set $\Lambda(q)$ of cost-sharing contracts acceptable to both agents is clearly non-empty, because lines L and T are parallel, with a common slope of $(1+r)$, and L is higher than T by a distance of $r > 0$.[9]

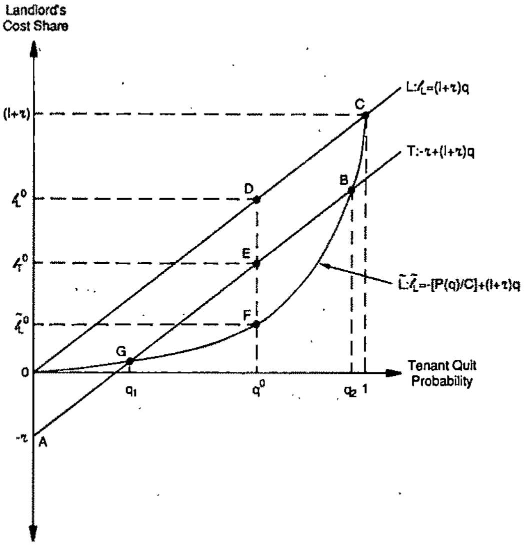


FIGURE 1

To see that with at least one risk-averse agent the set of mutually acceptable cost-sharing contracts may be empty, suppose the tenant is risk neutral but the landlord is risk averse. Then the tenant's behavior is still described by (2), (4), and line T in Figure 1. However, the value of the innovation to the landlord is now given by

$$(1') \quad \tilde{V}_L(l, q, r) = Eu(W)$$
$$= qu(W_0 + X - lC)$$
$$+ (1-q)u(W_0 - lC),$$

where W_0 is the landlord's initial wealth, $W = W_0 + x - lC$ is his terminal wealth, $x = [X, 0; q, 1-q]$ is the random variable capturing the output that accrues to the landlord, E is the expectation operator, and $u(\cdot)$ is the landlord's von Neuman Morgenstern utility function with $u'(\cdot) > 0$. Assume that $u''(\cdot) < 0$ in order to capture the land-

[8] Of course, it is to be understood that for a given quit probability $q = q^0$, the only mutually acceptable contracts are $\Lambda(q^0) = [l_T(q^0), l_L(q^0)]$.

[9] $l = q^0$ is one such contract. For it can be seen from (1) and (2) that for $l = q^0$, $V_L = q^0(X - C) > 0$ and $V_T = (1 - q^0)(X - C) > 0$.

lord's risk aversion. Further, since it is my purpose to demonstrate that *there are situations* where desirable innovations get rejected, I shall assume that the landlord's utility function is a constant absolute risk-aversion utility function,

$$u(W) = -ke^{-zW}, \qquad k, z > 0.$$

Since my purpose is to prove an *existence* result, restricting the analysis to a special class of utility functions is a harmless assumption.[10]

The landlord will accept all cost-sharing contracts as long as his expected utility with the innovation exceeds the utility of his initial wealth,

$$(7) \qquad Eu(W) \geq u(W_0).$$

From the definition of risk premium, we have

$$(8) \qquad u[\overline{W} - P(q)] = Eu(W),$$

where $\overline{W} = W_0 + qX - lC$ is the expected terminal wealth, and $P(q)$ is the risk premium associated with the gamble. In general the risk premium is a function of W_0, l, c, z, X, and q. However, as shown in the Appendix, under the constant absolute risk-aversion preferences, the risk premium is a function only of z, X, and q. Since I never consider a change in the output X, or in the magnitude of risk aversion z, it is legitimate to think of the risk premium as a function only of the tenant's quit probability q. Further, given an uncertain prospect, the risk premium $P(q)$ is strictly positive for a risk-averse agent and is higher for a more risk-averse agent.[11] From (7) and (8) it now follows that the landlord is willing to offer contracts for the adoption of the innovation for which $u[\overline{W} - P(q)] \geq u(W_0)$, or $\overline{W} - P(q) \geq W_0$, which, in turn, is tantamount to $qX - lC \geq P(q)$. That is, the set of all contracts the landlord finds desirable is

$$(3') \qquad \tilde{L} = \left\{ l \colon l \leq -\frac{P(q)}{C} + (1+r)q, \quad P(q) \geq 0, \quad r > 0, \quad 0 \leq q \leq 1 \right\}.$$

The equality in (3′) determines $\tilde{l}_L = -[P(q)/C] + (1+r)q$, which is the maximum cost share the landlord is willing to bear given the quit probability q. And this may well be smaller than the minimum landlord's cost share (see (4)) that the tenant is willing to accept [i.e., $l_T = -r + (1+r)q$]. This will indeed be the case if and only if $P(q)/C > r$ or, equivalently, $P(q) > (X - C)$. Therefore the set of mutually acceptable cost-sharing contracts is

$$(9) \quad \tilde{\Lambda} = \tilde{L} \cap \overline{T} = \phi \text{ if } P(q) > (X - C).$$

This conclusion is not at all unreasonable. It merely says that if the landlord is risk averse and if his risk premium exceeds the social profit of the innovation, then while he will offer the tenant innovation-cost-sharing contracts, none of these contracts will be good enough for the tenant, despite the innovation being such as to add more to property value than the cost of the innovation. All that remains to be proved now is that $P(q) > (X - C)$ is possible. And this condition is met by a C sufficiently close to X since, as shown in the Appendix, for the constant absolute risk aversion class of utility functions, P is not a function of C.

Additional insight is gained from Figure 1. The graph of $\tilde{l}_L(q) = -[P(q)/C] + (1 + r)q$ is presented in Figure 1 as the curve $\tilde{L}$. The particular shape of the curve $\tilde{L}$ derives from the properties of $P(q)$ established in the Appendix: $\tilde{l}_L(0) = 0$, $\tilde{l}_L(1) = (1 + r)$, $\tilde{l}'(q) > 0$ for all q, and $\tilde{l}_L(q)$ is strictly convex since $P(q)$ is strictly concave and $l''_L(q) = -P''(q)/C$. For a quit probability such as q^0, the tenant accepts all contracts $l \geq l^0_T$, while the landlord is only willing to offer contracts $l \leq \tilde{l}^0_L$, so that for this quit probability there exists no mutually beneficial in-

[10] This was pointed out to me by the referee, who also brought my attention to the fact that the constant absolute risk-aversion utility function has the added feature that it makes the risk premium independent of l, a feature that permits a very simple proof of the central result of this paper.

[11] See Kenneth Arrow (1965) and John Pratt (1964).

novation-cost-sharing contract. In fact, with the help of Figure 1, a stronger assertion can be made. For every (X,C), if for some tenant quit probability q it is the case that $P(q)>(X-C)$, then there exists an interval (q_1,q_2) such that the innovation is rejected for all $q\in(q_1,q_2)$, while it is accepted for all sufficiently large or sufficiently small values of $q\in[0,q_1]\cup[q_2,1]$. This conclusion clearly demonstrates that there are circumstances in which, under the tenurial laws, both a landlord with a single-tenant-occupied rental property and the tenant will find that it is not worth their while to innovate, so that suboptimal investment and innovation, and hence technological stagnation, will be the outcome.[12] It may also be noted that, since for a given gamble the risk premium is higher for a more risk-averse agent, it will be the destiny of the relatively less risk-averse landlords and their tenants to enjoy the fruits of innovation, while, under the same tenurial laws, the relatively more risk-averse landlords and their tenants will not reach any mutually beneficial agreements regarding investment in property improvement.[13]

II. Conclusion

There are several reinforcing reasons for technological stagnation in less developed countries. Some of these reasons are associated with the peculiar nature of the tenurial laws in these countries, and Kaushik Basu (1989) has successfully explained technological stagnation as an outcome of the tenurial laws in the case of those landlords who own several properties. A large proportion of landlords in these countries, however, own only a single property. I have explained technological stagnation as an outcome of the tenurial laws in the single-property-holding case based on the work on decision making under uncertainty by Kenneth Arrow (1965) and John Pratt (1964); the argument I have developed is equally valid for the multiple-property-holdings case and, in that case, further reinforces Basu's adverse selection argument for stagnation.

Not only do several less developed countries suffer from technological stagnation, but there seem to be institutions and laws in place that tend to perpetuate this stagnation. Tenurial law amendments in favor of those who have the capacity to innovate are called for; relatively minor amendments may well foster a climate of tremendous technological progress.

APPENDIX

In this appendix I establish those properties of the risk-premium function that are used in the proof of the central result of this paper and that underlie the construction of Figure 1. The notation here is the same as in the text.

Under the constant absolute risk-aversion utility function

$$\text{(A1)}\qquad u(W)=-ke^{-zW},\qquad k,z>0,$$

from (8) we have

$$\text{(A2)}\quad -ke^{-z[W_0+qX-lC-P]}$$
$$=-k\left[qe^{-z(W_0+X-lC)}+(1-q)e^{-z(W_0-lC)}\right].$$

In (A2) the LHS is $u(\overline{W}-P)$ and the RHS is

$$Eu(W)=qu(W_0+X-lC)+(1-q)u(W_0-lC).$$

From (A2) we have

$$\text{(A3)}\qquad e^{-z(qX-P)}=qe^{-zX}+(1+q).$$

From (A3) it can be concluded that the risk premium P is a function only of z, X, and q and not a function of W_0, l, or C. (A3) also permits an explicit solution of P. Taking the natural logarithms of (A3), straightforward manipulation yields

$$\text{(A4)}\quad P(q,X,z)=\frac{1}{z}\ln\left[qe^{-zX}+(1-q)\right]+qX.$$

Since in this paper the magnitude of risk aversion z and the output X are treated as constants, let $z=X=$

[12] While in this paper I have developed the argument for the case of a risk-averse landlord and a risk-neutral tenant, an identical argument is valid if the landlord is risk neutral and the tenant risk averse.

[13] If landlord A is more risk averse than landlord B, then $P_A(q)>P_B(q)$. If $P_A(q)/C>r$, so that A and his tenant do not adopt the innovation, then for a sufficiently less risk-averse B, $P_B(q)/C<r$, so that B and his tenant decide to jointly finance the innovation.

1. Then from (A4) we have

$$P(q)=\ln[(1-q)+q/e]+q. \tag{A5}$$

From (A5) it may first be noted that $P(0)=\ln(1)=0$ and that $P(1)=\ln(e^{-1})+1=0$. Further,

$$P'(q)=\frac{(e^{-1}-1)}{[(1-q)+q/e]}+1 \tag{A6}$$

and

$$P''(q)=-\left\{\frac{(e^{-1}-1)}{[(1-q)+q/e]}\right\}^2<0. \tag{A7}$$

(A7) clearly demonstrates that the risk premium is a strictly concave function of q. And from (A6) we have $P'(0)=1/e>0$, $P'(1)=2-e<0$, and $P'[1/(e-1)]=0$. Hence $P(q)$ starts out equal to zero at $q=0$, increases initially as q rises, then decreases, and finally equals zero again at $q=1$.

REFERENCES

Arrow, Kenneth J., *Aspects of the Theory of Risk Bearing*, Helsinki: Johnsonin Sattie, 1965.

Bardhan, Pranab K., *Land, Labor and Rural Poverty: Essays in Development Economics*, New York: Columbia University Press, 1984.

Basu, Kaushik, "Technological Stagnation, Tenurial Laws, and Adverse Selection," *American Economic Review*, March 1989, *79*, 251–55.

Bhaduri, Amit, "A Study in Agricultural Backwardness Under Semi-Feudalism," *Economic Journal*, March 1973, *83*, 120–37.

Braverman, Avishay and Stiglitz, Joseph, "Landlords, Tenants and Technological Innovations," *Journal of Development Economics*, October 1986, *23*, 313–32.

Johnson, D. Gale, "Resource Allocation Under Share Contracts," *Journal of Political Economy*, April 1950, *58*, 111–23.

Kalecki, Michal, *Essays on Developing Economies*, Hassocks: Harvester Press, 1976.

Newbery, David M. G., "Tenurial Obstacles to Innovation," *Journal of Development Studies*, July 1975, *11*, 263–77.

Pratt, John W., "Risk Aversion in the Small and in the Large," *Econometrica*, January–April 1964, *32*, 122–36.

Quality Distortion by a Discriminating Monopolist: Comment

By SHABTAI DONNENFELD AND LAWRENCE J. WHITE*

> Assume that a monopolist makes men's suits, and that he makes only one size of suit. This is absurd behavior, but the picture of the sadistic monopolist who disregards consumer desires has often made fugitive appearances in the literature....
>
> The general conclusion that I wish to draw is that a monopolist who does not cater to the diversities of his buyers' desires will suffer a substantial decline in his profits.
>
> [George J. Stigler, 1964, pp. 60–61]

A number of recent theoretical papers have shown Stigler's conjecture to be incorrect.[1] A monopolist that offers multiple product quality levels to multiple consumer types may well find it worthwhile to distort some quality levels or not provide them at all. This distortion can enhance the monopolist's profits because it prevents or discourages some customer types, who have a high willingness to pay for some quality levels, from switching to other quality levels that yield lower profits for the monopolist.

In a recent paper, Padmanabhan Srinagesh and Ralph M. Bradburd (1989) show that this quality distortion can involve higher-than-optimal quality levels as well as quality deterioration. This finding closely parallels the results previously demonstrated by Shabtai Donnenfeld and Lawrence J. White (1984, 1988).

The Srinagesh-Bradburd discussion of their model neglects any efforts at comparative statics. We believe that these comparative statics are interesting; they indicate, for example, that the comparative *numbers* of each customer type are an important influence on the degree of quality distortion that occurs and on the prices that the monopolist charges for each quality level. The remainder of this paper provides and discusses these comparative statics results as well as some policy-related issues.

I. A Review of the Model

The Srinagesh-Bradburd and Donnenfeld-White models are structured as follows.[2]

A consumer is interested in buying (at most) only one unit of a specific product that can be produced in a continuous array of qualities that are indexed by Q, with higher Q implying higher quality; the remainder of the consumer's consumption is of a composite good, X. The consumer pays a price P for the unit of the good of quality Q. The consumer's indirect utility function can be represented by $U(Q,P)$, and the lowest level of utility to which he can be forced (at which point he will consume only X) is $U(0,0)$.[3]

There are two types of consumers, a and b, numbering N_a and N_b, respectively. For all (Q,P) we assume that $U_a(Q,P) < U_b(Q,P)$; that is, type b consumers have a higher absolute willingness to pay for quality.

The product is produced by a monopolist at a unit cost of $c(Q)$, with $c'(Q) > 0$ and $c''(Q) \geq 0$. The monopolist knows the utility functions of the two types of consumers,

*Stern School of Business, New York University, and Graduate School of Business, Tel Aviv University; and Stern School of Business, New York University, New York, NY 10006.

[1] These papers include Lawrence White (1977); Michael Mussa and Sherwin Rosen (1978); Eric Maskin and John Riley (1984); Shabtai Donnenfeld and Lawrence White (1984, 1988); J. J. Gabszewicz et al. (1986); David Besanko, Shabtai Donnenfeld, and Lawrence J. White (1987, 1988); and Padmanabhan Srinagesh and Ralph M. Bradburd (1989).

[2] To ease the comparability of our results with the Srinagesh-Bradburd model, we adopt their notation and formulation rather than that used in our earlier model (Donnenfeld and White, 1984, 1988).

[3] This specification of the utility function assumes away income effects.

and he knows their numbers; but he does not know their identities, and he cannot keep them separated through any physical or legal means; that is, each consumer type is able to buy any product quality that is offered on the market at any prevailing price.

The monopolist's problem, then, is to maximize

$$(1)\qquad \Pi = N_a\{P_a - c(Q_a)\} + N_b\{P_b - c(Q_b)\}$$

subject to

$$(2)\qquad U_a(Q_a, P_a) - U_a(Q_b, P_b) \geq 0$$

$$(3)\qquad U_b(Q_b, P_b) - U_b(Q_a, P_a) \geq 0$$

$$(4)\qquad U_a(Q_a, P_a) - U_a(0,0) \geq 0$$

$$(5)\qquad U_b(Q_b, P_b) - U_b(0,0) \geq 0.$$

If the monopolist chooses to supply goods to both consumer types, the only binding constraints are (3) and (4), that is, the incentive compatibility constraint (ICC) and the individual rationality constraint (IRC), respectively. The monopolist's profit maximization problem can be written as

$$\underset{(Q_a, Q_b)}{\text{Max}}\ \Pi = N_a\{U_a(Q_a) - c(Q_a)\} + N_b\{U_b(Q_b) - U_b(Q_a) + U_a(Q_a) - c(Q_b)\}.$$

The necessary and sufficient conditions are

$$(6)\qquad \Pi_{Q_a}(Q_a(N_a, N_b, \theta_a, \theta_b, \gamma_a, \gamma_b, \phi), Q_b(N_b, \theta_b, \gamma_b, \phi), N_a, N_b, \gamma_a, \gamma_b, \phi) = 0$$

$$(7)\qquad \Pi_{Q_b}(Q_b(N_b, \theta_b, \gamma_b, \phi), N_b, \phi) = 0.$$

In (6) and (7) we explicitly depict the dependence of the endogenous quality variables on parameters that shift the utility and cost functions; θ_i provides a constant increment to U_i, and γ_i affects only the marginal willingness to pay for quality. Similarly, $c = c_0 + \phi c(Q)$. Note that Q_a depends on all the exogenous parameters, while Q_b depends only on the parameters directly related to type b consumers.

As Srinagesh and Bradburd (1989) and Donnenfeld and White (1984, 1988) demonstrate, the solution to this problem depends on the marginal valuation of quality (*MVQ*) of the type a and type b consumers. If, for all Q, $MVQ_a < MVQ_b$—that is, if absolute and marginal willingness to pay by the two consumer types are positively correlated (the "standard" assumption)—then the monopolist maximizes profits by distorting downward the quality level of the lower-quality product.[4] If, however, $MVQ_a > MVQ_b$—if absolute and marginal willingness to pay for quality are negatively correlated (the "nonstandard" assumption)—then quality distortion occurs through an *increase* in the quality level of the higher-quality product.[5]

One other point is worth noting. As White (1977) and Donnenfeld and White (1984, 1988) demonstrate and as Srinagesh and Bradburd (1989, fn. 5) note, the incentive for the monopolist to distort quality may be strong enough that he reaches a boundary solution in which only a single quality level is produced for the monopolist's dominant (i.e., most profitable) customers; the customers who were potentially subject to the quality distortion find that they are not provided with any suitable quality level whatsoever.[6]

[4] Quality is distorted in the sense that the quality level provided to group a is different from the quality level that a competitive market would provide to group a or that a monopolist that could keep the two groups apart would provide to group a.

[5] See Donnenfeld and White (1984, 1988) for some economic interpretations of this assumption.

[6] See Stephen W. Salant (1989) and Hal R. Varian (1989) for a unifying formal treatment of the necessary and sufficient conditions that induce internal or corner solutions in atemporal and intertemporal models of price discrimination via self-selection.

TABLE 1—THE COMPARATIVE STATICS OF THE "STANDARD" CASE, IN WHICH THE LOW-QUALITY PRODUCT IS DISTORTED DOWNWARD (POSITIVE CORRELATION OF ABSOLUTE AND MARGINAL WILLINGNESS TO PAY)

	dN_a	dN_b	$d\theta_a$	$d\theta_b$	$d\gamma_a$	$d\gamma_b$	$d\phi$
dQ_a	+	−	0	0	+	−	−
dP_a	+	−	+	0	+	−	−
dQ_b	0	0	0	0	0	+	−
dP_b	−	+	+	0	−	+	−
Likelihood of Specializing on the High-Quality Product	−	+	−	+	−	+	−

TABLE 2—THE COMPARATIVE STATICS OF THE "NONSTANDARD" CASE, IN WHICH THE HIGH-QUALITY PRODUCT IS DISTORTED UPWARD (NEGATIVE CORRELATION OF ABSOLUTE AND MARGINAL WILLINGNESS TO PAY)

	dN_a	dN_b	$d\theta_a$	$d\theta_b$	$d\gamma_a$	$d\gamma_b$	$d\phi$
dQ_a	−	+	0	0	+	+	−
dP_a	−	+	+	0	+	+	−
dQ_b	0	0	0	0	0	+	−
dP_b	−	+	+	0	+	+	−
Likelihood of Specializing on the High-Quality Product	−	+	−	+	−	+	−

II. Comparative Statics

The comparative statics for the "standard" (positive correlation) and "nonstandard" (negative correlation) cases are provided in Tables 1 and 2, respectively.

For illustrative purposes, we shall provide the explicit derivation of the effects of changes in the number of consumers for the standard case—positive correlation between absolute and marginal willingness to pay for quality. This derivation explains how we arrive at the first two columns in Table 1. Differentiation of (6) and (7) with respect to N_a and N_b, respectively, yields

$$(8)\qquad \frac{\partial Q_a}{\partial N_a} = -\frac{\Pi_{Q_aN_a}}{\Pi_{Q_aQ_a}} > 0$$

$$\frac{\partial Q_a}{\partial N_b} = -\frac{\Pi_{Q_aN_b}}{\Pi_{Q_aQ_a}} < 0$$

$$(9)\qquad \frac{\partial Q_b}{\partial N_a} = \frac{\partial Q_b}{\partial N_b} = 0,$$

where $\Pi_{Q_iQ_i} < 0$ by concavity of the profit function and $\Pi_{Q_aN_a} = U_a'(Q_a) - C'(Q_a)$ is positive by the first-order condition and $\Pi_{Q_aN_b} = -[U_b'(Q_a) - U_a'(Q_a)]$ is negative by the standard assumption.

Further, differentiation of the ICC (3) and IRC (4) with respect to N_a and N_b and using (8) and (9) yields

$$(10)\qquad \frac{\partial P_a}{\partial N_a} = U_a'(Q_a)\frac{\partial Q_a}{\partial N_a} > 0,$$

$$(11)\qquad \frac{\partial P_b}{\partial N_a} = U_b'(Q_b)\frac{\partial Q_b}{\partial N_a} - [U_b'(Q_a) - U_a'(Q_a)]\frac{\partial Q_a}{\partial N_a} = -[U_b'(Q_a) - U_a'(Q_a)]\frac{\partial Q_a}{\partial N_a} < 0,$$

$$(12)\qquad \frac{\partial P_a}{\partial N_b} = U_a'(Q_a)\frac{\partial Q_a}{\partial N_b} < 0, \text{ and}$$

$$(13)\qquad \frac{\partial P_b}{\partial N_b} = U_b'(Q_b)\frac{\partial Q_b}{\partial N_b} - [U_b'(Q_a) - U_a'(Q_a)]\frac{\partial Q_a}{\partial N_b} = [U_b'(Q_a) - U_a'(Q_a)]\frac{\partial Q_a}{\partial N_b} > 0.$$

Some interesting results emerge from these comparative statics.

(i) First, the relative numbers of the two consumer types matter. As the relative number of the high absolute willingness to pay consumers (N_b) increases, the quality distortion experienced by the other group worsens. In the standard case—positive correlation—the quality provided to type a consumers deteriorates, and its associated price declines (see Table 1). In the nonstandard case—negative correlation—the quality provided to type a consumers is increased, and the price they pay rises (see Table 2). Furthermore, in both cases it is more likely that the monopolist will specialize in producing products only for those in group b as their number increases. The intuition behind these results is straightforward: As the relative number of the high absolute willingness to pay (high profit) customers (N_b) increases, the monopolist can increase his overall profits by worsening the alternative (Q_a). He thereby loses some profits in his sales to group a, but he gains even larger profits by being able to raise his price (P_b) to group b.

(ii) Second, changes in either group's absolute willingness to pay have no effects on quality distortion, though they do affect the location of the boundary solution and therefore affect the likelihood that the monopolist will specialize on producing products only for the b group. Also, an increase in group a's absolute willingness to pay will mean an increase in the prices that both groups pay. The intuition here is that an increase in a's absolute willingness to pay for quality means that the monopolist can charge a higher price to this group; this higher price means, in turn, that the (Q_a, P_a) alternative is still less attractive to group b, so the monopolist can now charge a higher price (P_b) to that group.

(iii) Finally, an increase in the marginal willingness to pay by group b in the standard case causes the quality distortion (deterioration) experienced by group a to worsen and increases the likelihood that the monopolist will specialize on group b. The price charged to group b increases, partly because of the pure effect of b's increased marginal willingness to pay and partly because the monopolist has worsened the alternative (Table 1). However, in the nonstandard case, an increase in the marginal willingness to pay by group b results in further distortion, albeit a too high a quality is provided to these consumers. Group a buyers end up paying a higher price for the same quality (Table 2).

With a suitable reinterpretation of the variables presented in Tables 1 and 2, we can use these comparative statics briefly to examine a few policy issues. For example, the imposition of a production tax is the equivalent of an increase in the cost parameter ϕ. Hence, for both the standard and nonstandard cases, levying this tax of $d\phi$ on a multiproduct monopolist results in an overall decline in quality *and prices* of the goods provided to all consumers.

Further, the imposition of a binding import quota on goods of a specific quality type supplied by a multiproduct foreign monopolist is the equivalent of a decrease in the number of consumers for that quality type. Hence, for the standard case, a quota on low-quality (high-quality) imports results in the monopolist's downgrading (upgrading) of the lower-quality imports; for the nonstandard case, a quota on high-quality (low-quality) imports results in the upgrading (downgrading) of the higher-quality imports.[7]

[7] See Shabtai Donnenfeld (1988) for a more elaborate discussion on trade policies.

III. Conclusions

Microeconomic models involving endogenous quality continue to provide a rich vein for theoretical research and policy-oriented issues. As we have shown, the comparative statics analysis of the Srinagesh-Bradburd (1989) and Donnenfeld-White (1984, 1988) models adds to the insights that can be gained from these models.

REFERENCES

Besanko, David, Donnenfeld, Shabtai and White, Lawrence J., "Monopoly and Quality Distortion: Effects and Remedies," *Quarterly Journal of Economics*, November 1987, *102*, 743–68.

______, ______ **and** ______, "The Multiproduct Firm, Quality Choice, and Regulation," *Journal of Industrial Economics*, June 1988, *36*, 411–30.

Donnenfeld, Shabtai, "Commercial Policy and Imperfect Discrimination by a Foreign Monopolist," *International Economic Review*, November 1988, *29*, 607–20.

______ **and White, Lawrence J.,** "Product Variety and the Inefficiency of Monopoly: A Simple Model and an Examination of Some Possible Remedies," New York University, Graduate School of Business Administration, Working Paper No. 84–82, October 1984.

______ **and** ______, "Product Variety and the Inefficiency of Monopoly," *Economica*, August 1988, *55*, 393–401.

Gabszewicz, J. J., Shaked, A., Sutton, J. and Thisse, J. F., "Segmenting the Market: The Monopolist's Optimal Product Mix," *Journal of Economic Theory*, August, 1986, *39*, 273–89.

Maskin, Eric and Riley, John, "Monopoly with Incomplete Information," *Rand Journal of Economics*, Summer 1984, *15*, 171–96.

Mussa, Michael and Rosen, Sherwin, "Monopoly and Product Quality," *Journal of Economic Theory*, August 1978, *18*, 301–17.

Salant, Stephen W., "When Is Inducing Self-Selection Suboptimal for a Monopolist?" *Quarterly Journal of Economics*, May 1989, *104*, 391–98.

Srinagesh, Padmanabhan and Ralph M. Bradburd, "Quality Distortion by a Discriminating Monopolist," *American Economic Review*, March 1989, *79*, 96–105.

Stigler, George J., "A Theory of Oligopoly," *Journal of Political Economy*, February 1964, *72*, 44–61.

Varian, Hal R., "Price Discrimination," in R. Schmalensee and R. Willig, eds., *Handbook of Industrial Organization*, Amsterdam: North-Holland, 1989.

White, Lawrence J., "Market Structure and Product Varieties," *American Economic Review*, March 1977, *67*, 179–82.

Comparative Productivity: Comment

By STEVEN ROSEFIELDE*

Abram Bergson (1987*a*) has recently attempted to show that socialist economic systems are underproductive and underefficient.[1] Close evaluation, however, reveals that this conclusion is too sweeping. Using the USSR as a case in point, this comment demonstrates that Soviet productivity growth using Bergson's data (1987b) and methods (Bergson, 1978) matched Western Europe's and exceeded the U.S. achievement in 1960–75. Moreover the pronounced duality between Soviet industrial and nonindustrial productivity discovered by Bergson for 1960 appears to have increased thereafter through 1975, according to the statistics of the United States Central Intelligence Agency (CIA) similar to those employed in Bergson (1987b). This duality is confirmed further by a separate calculation of comparative machine-building productivity. Using CIA sector of origin data that are consistent with Bergson's counterpart end use data (Imogene Edwards et al., 1979) and his factor share weights (Bergson, 1978), Soviet machine-building productivity is calculated to be 79 percent of the American level in 1975 (Table 1). Revised data for the same year, published by the CIA in 1982, raise this figure to 88 percent. These findings clearly suggest that socialist economies need not be comprehensively underproductive (underefficient). Although they do not dispose of many issues of legitimate controversy[2] and pose serious questions about the potential shortcomings of the adjusted data on which Bergson relies, they do demonstrate nonetheless that the positivist debate over the comparative productivity (efficiency) of socialism even in the Soviet case still cannot be completely laid to rest.[3]

I. Underproductivity

Specialists have never doubted that the aggregate productivity of the American economy surpasses the Soviets'. A clear picture of the dimensions of the U.S. advantage, given the inscrutability of Soviet statistics, however, did not emerge until the late 1960's when Bergson published *Planning and Productivity Under Soviet Socialism*. His principal findings, as perfected in his later works, are reproduced in Table 2. They are expressed as a series of "coefficients of factor productivity" for 1960 that compare factor productivity in Western Europe and the USSR with the United States on the explicit assumption that all countries possess a common linear homogeneous, Cobb-Douglas, transnational production function.[4] Output is measured by gross domestic material

*The author is at the University of North Carolina, Campus Box 3305, Gardner Hall, Chapel-Hill, NC 27599-3305. The author wishes to thank three anonymous referees for their constructive suggestions and the Earhart Foundation for its support.

[1] Bergson, 1987a, p. 355. (Compare Bergson, 1948, pp. 235–236.) For a discussion of the distinction between underproductivity and underefficiency, see Bergson (1978, pp. 68–74).

[2] They do not take full account of relative backwardness. Presumably, Soviet productivity growth should have exceeded, rather than matched the West European experience. They ignore welfare losses caused by the unresponsiveness of supply to consumer demand. And it can be argued, contrary to Bergson's judgment, that they seriously underestimate the effects of hidden inflation (Bergson, 1987c, p. 420).

[3] Bergson carefully elaborates the theoretical and empirical limitations of his empirical calculations, including issues of technology, economies of scale, development stage, milieu, and degrees of freedom. He acknowledges that his findings are not definitive, but argues for their significance by observing that "Such a qualified potentiality, however, is not what proponents of socialism have usually claimed for that system" (Bergson, 1987a, p. 356). Differences in the quality of land, labor, and capital that are not captured in the aggregate input data should also be added to Bergson's inventory of shortcomings.

[4] The appendix to this paper (available upon request) discusses coefficients of factor productivity.

TABLE 1—COMPARATIVE INTERNATIONAL AGGREGATE, INDUSTRIAL, AND MACHINE-BUILDING EFFICIENCY IN 1960 AND 1975

	Summary Estimates: Coefficients of Factor Productivity (Valued at U.S. Dollar Prices; Index U.S. = 100)			
	GNP		Industry	MBMW
Country	1960	1975	1960	1975
United States	100	100	100	100
Western Europe	60	85	65	–
USSR	41	55	58	79

Sources: Tables 2, 3, and A12 (in statistical appendix, available upon request). *Method:* Coefficients of factor productivity following Bergson are computed according to the formula

$$\theta = (y_i/y_j)/\left[(k_i/k_j)^{\alpha}(v_i/v_j)^{\beta}(l_i/l_j)^{1-\alpha-\beta}\right],$$

where the ith subscript refers to the Soviet Union or Europe and the jth to the United States; y, k, v and l are output, fixed capital, inventories, and labor, and the exponents are factor shares based on American dollar weights. Inventories are omitted in computing coefficients of MBMW productivity. θ values for Western Europe are simple averages of the figures for France, Germany, Italy, and the UK. GNP is defined as gross material product, excluding selected final services: health care, education, government administration, defense, and housing. Industry is defined broadly and includes manufacturing, mining, power, construction, transport and communications, and trade. Statistics in columns 1 and 3 are Bergson's; those in column 2 are derived directly from Bergson's data; those in column 4 are Rosefielde's based on the CIA's estimates of Soviet MBMW output and growth provided in CIA (1975) and Edwards et al. (1979, p. 273). See sources and the appendix for further details.
Notes: The MBMW statistics reported above have not been adjusted for hidden inflation. For alternative estimates see the text, fn. 13, and Tables A12 and A16 in the statistical appendix (available upon request). Weapons constituted 36 percent of Soviet MBMW at established ruble prices in 1970.

product (except in Soviet/American comparisons, where gross national material product is employed) exclusive of output originating in selected final services: health care, education, government administration, defense, and housing. Nonfarm labor is expressed in adjusted man-hours. Both labor and reproducible fixed capital have the same coverage as output.

The coefficients of factor productivity reveal that both labor and combined factor productivity were significantly lower in the Soviet Union than in the West. Given the uncertainties that beset international comparisons, this finding is not conclusive, but it is persuasive because the principal norm adopted favors the Soviet Union.[5] Soviet factor productivity is computed in dollars, a procedure that, as is widely understood, overstates its relative position.

In 1960 Soviet labor productivity was only 31 percent of the U.S. level, and combined

[5] The ruble data Bergson uses correspond closely with official Soviet statistics and are not significantly adjusted for hidden inflation. According to Igor Birman (1983) the CIA dollar-ruble ratios Bergson employs overstate the dollar value of Soviet output. The CIA itself has revised the estimates on which Bergson's analysis depends, sharply downward (CIA, 1984). It follows directly therefore that insofar as Bergson's estimates err, they probably do so in the Soviet's favor. (Compare Rosefielde, 1988a.)

TABLE 2—COEFFICIENTS OF FACTOR PRODUCTIVITY
(U.S. = 100)

Country	Gross Material Product per Employed Worker		Gross Material Product per Unit of Factor (Labor and Reproducible Capital) Inputs	
	1960	1975	1960	1975
United States	100	100	100	100
France	51	94	63	96
Germany	51	91	65	92
United Kingdom	49	69	64	73
Italy	34	71	47	77
USSR	31	58	41	55

Sources: Tables A1 and A2 (available upon request) and Bergson (1978: table 6.1, pp. 76–77; table 7.1, p. 93; appendix tables 11, p. 236, and 18, p. 241).

Data: Gross material product represents gross domestic product exclusive of output originating in selected final services: health care, education, government administration, defense, and housing. Dollar estimates are valued at prevailing 1960 and 1975 U.S. prices. U.S. and Soviet data are consistently defined. Major inputs refer to labor and reproducible fixed capital. Employment is adjusted for differences in nonfarm hours; capital is calculated as the average of two relatives, one with fixed capital including gross of depreciation and the other with such capital including net of depreciation.

Method: The coefficients of factor productivity for 1960 are reproduced directly from Bergson (1978). The coefficients of factor productivity for 1975 are computed by the author, using the data provided in Bergson's unpublished statistical appendix (Bergson 1987b), according to the formula

$$\theta=(y_i/y_j)/\left[k_i/k_j)^{\alpha}(v_i/v_j)^{\beta}(l_i/l_j)^{1-\alpha-\beta}\right],$$

where the ith subscript refers to the Soviet Union, or a specific European country, and the jth to the United States; y, k, v, and l are output, fixed capital, inventories, and labor, and the exponents are factor shares based on American dollar weights.

Labor, reproducible fixed capital inputs, and inventories are aggregated logarithmically with weights of 0.7415, 0.2278, and 0.0307, respectively, reported by Bergson for American factor income shares in 1960 (see Bergson, 1978, pp. 71, 95 fn. 3, and appendix table 18, p. 241). Bergson (1978) includes inventories in his calculations for 1960 (see appendix table 12, p. 238).

labor and capital productivity was 41 percent. This poor showing may have diverse technical causes, including production function peculiarities, unfavorable factor proportions, and economies/diseconomies of scale. But as Bergson (1978) himself notes, to the extent that production functions in the countries sampled approximate the linear homogeneous, transnational Cobb-Douglas standard, the low Soviet productivity (efficiency) appears to be primarily attributable to the shortcomings of its economic system.

Bergson's most recent contribution is aimed at further illuminating this important issue. He applies the same conceptual apparatus and data conventions to assess comparative factor productivity for a later year, 1975, but alters his approach in two ways. He expands his sample to include additional capitalist and socialist countries and estimates the systems underproductivity (underefficiency) effect directly, using various econometric techniques. His regressions indicate that socialist countries are underproductive and/or underefficient, and that the Soviet Union is the least efficient socialist country followed by Hungary, Poland, and Yugoslavia.

As with most econometric research, it is easy to fault these results. The assumption of a transnational Cobb-Douglas production function is implausible.[6] The sample size is small. With the exception of Spain, the levels of development between socialist and capitalist countries diverge noticeably. The unifying factors that distinguish Soviet, Polish, Hungarian, and Yugoslavian socialism from socialist democratic arrangements in West Europe and presumably cause their underefficiency are neither clearly explained nor convincingly defended. And the underlying conversion of Edwards' ruble estimates with the CIA's dollar-ruble ratios is highly suspect.[7] Nonetheless, on the basis of Bergson's prior studies, it seems unlikely that reasonable allowance for these factors would significantly alter his principal conclusion. Judged by gross material product, the Soviet Union and its fellow socialist economies of the East are underproductive and probably underefficient as well after proper adjustment is made for levels of economic development and sundry other considerations.

This aggregate assessment is important, but its scope is more restrictive than it seems. Bergson's econometric calculations pertain exclusively to the year 1975 and fail to illuminate whether the socialist economies of the East are overcoming their relative international underproductivity (underefficiency). Also, they are likely to conceal sectoral asymmetries. The Russian economy has manifested a distinct duality since the late nineteenth century. Industry in general and heavy industry in particular have been markedly more productive than other components of gross material product, and it would be surprising if this has ceased to be the case. Although the Soviet system as a whole almost certainly is underefficient, it might still be successful in some important activities.

Evidence supporting the first hypothesis is provided in Table 2. It presents coefficients of factor productivity computed with data provided in Bergson (1987b), according to Bergson's original methodology,[8] measured with respect to employment and other combined factors for the years 1960 and 1975. The improvement in relative Soviet factor productivity (efficiency) during this period is remarkable. Relative gross material product per employed worker is 87 percent higher in 1975 than in 1960. The gain in relative gross material product per unit of factor (labor and reproducible capital) inputs is less pronounced but still an impressive 34 percent. The magnitudes of these advances are in line with the European mean, which is less than might have been expected given the Soviet Union's relative backwardness.[9] This shortcoming not withstanding, however, the USSR did reduce its overall relative underproductivity (underefficiency) vis-à-vis the West.

The duality of the Soviet economy also manifests itself in Bergson's data. Table 3 reproduces his estimates of gross industrial product per employed worker and per unit of factor inputs in selected countries in 1960. They are calculated as before but omit agriculture as well as selected services. Relative Soviet factor productivity is clearly higher in industry than in agriculture. Instead of conspicuously lagging the pack, Soviet industrial factor productivity was more or less on a par with Italy and the United Kingdom

[6] The close correspondence between the stage of Soviet and Italian development and their aggregate productivity in 1960 had suggested that the transnational production function was understating Soviet inefficiency. This is no longer the case. Compare their coefficients of factor productivity for 1975 in Table 2.

[7] See fn. 5, Bergson (1987a, p. 347), and Bergson (1987b, table 3, p. 3).

[8] Bergson's econometric estimates could not be applied to his 1960 data set because the required land statistics are unavailable.

[9] Bergson (1987a) shows that Soviet static factor productivity is less than it should be after adjustment for the level of economic development. The dynamic results are mixed, although even here it can be argued that Bergson's results are confirmed with respect to Italy and, perhaps, England as well. See Table 2. As a consequence they qualify Bergson's finding that the Soviet Union is always underefficient, adjusted for the level of development. Bergson (1978) addresses the subject of dynamic productivity in his table 9.4 (p. 158). His results, which cover the period 1955–70, are similar to those reported here. Soviet productivity growth 1960–75, however, was better relative to its Western rivals than in the time period Bergson studied.

TABLE 3—GROSS INDUSTRIAL PRODUCT PER EMPLOYED WORKER AND PER UNIT OF FACTOR INPUTS, SELECTED COUNTRIES, 1960

Country	Gross Industrial Product per Employed Worker	Gross Industrial Product per Unit of Factor (Labor and Reproducible Capital) Inputs
United States	100	100
France	60	71
Germany	54	69
United Kingdom	48	61
Italy	46	60
USSR	50	58

Source: Bergson (1978, table 7.5, p. 108).
Data: Gross industrial product represents essentially the gross output originating in manufacturing, mining, power, construction, transport and communications, and trade. In the calculation of output per worker and per composite unit of factor inputs, reference is to employment and capital stock used in the same sector. Valuation of output and inputs is in 1960 U.S. dollars. Employment is also adjusted for variations in the length of the workweek. Combined inputs are aggregated logarithmically according to the Cobb-Douglas specification. The method for computing the coefficient of factor productivity (column 2) is specified in Table 2.

and not glaringly below France and Germany.

Bergson has not published similar statistics for 1975, but official Soviet and CIA data can be used to show that the sectoral duality of the Soviet economy persists and has probably intensified. Table 4 reports estimates of aggregate and industrial factor productivity growth 1965–85. The output series are CIA estimates, as are the statistics on employment and capital. Industry is defined exclusive of construction, transport, communications, and trade. Inputs are combined according to the Cobb-Douglas specification with Bergson's income share weights. Industrial total factor productivity measured in this way exceeds the aggregate rate for the entire period 1965–85 and never falls below it for each quinquennial subperiod. The same comparison computed by using official industrial growth data suggests a more pronounced differential that exceeds 60 percent (Table A17 in statistical appendix, available upon request). Although analogous estimates for the United States and the West are not at hand, this disparity suggests that the Soviets not only significantly diminished their relative aggregate underproductivity, but may have reduced their industrial underproductivity as well.[10]

Industry is not a homogeneous activity. The Soviets have long accorded higher priority to producer durables than nondurable and durable consumer goods. This preference is reflected in the rapid growth of producer durables,[11] and in the relatively high level of machine-building productivity displayed in 1975 computed with the aid of CIA sector of origin statistics that are consistent with the counterpart end use data that Bergson employs and the methods he utilized in his earlier studies. This figure reported in Table 1 is 79 percent of the American standard.[12] An alternative estimate using revised CIA sector of origin statistics for the same year raises the figure to 88 percent, while other estimates based

[10] Compare Bergson, 1978, table 9.5, p. 158.

[11] See Bergson, 1987c.

[12] If the disparity between industrial and aggregate productivity reported by Bergson for 1960 (Table 1) is extrapolated to 1975, industrial productivity would be 78 percent of the U.S. level.

Table 4—Estimates of Soviet Factor Productivity Growth 1965–85

	CIA Statistics/Bergson's Weights				
Measure	1965–70	1970–75	1975–80	1980–85	1965–85
Gross National Product	4.9	3.1	2.3	2.2	3.1
Combined Inputs	3.8	3.8	3.1	2.5	3.3
Work Hours	2.0	1.7	1.2	0.7	1.4
Capital	7.4	8.0	6.9	6.3	7.1
Total Factor Productivity	1.1	−0.7	−0.8	−0.3	−0.2
Industrial Output	6.0	5.7	2.7	2.3	4.2
Combined Inputs	4.9	7.8	3.4	2.6	3.6
Work Hours	3.1	1.5	1.4	0.5	1.6
Capital	8.8	8.7	7.7	7.0	8.0
Total Factor Productivity	1.1	1.9	−0.7	−0.3	0.6
Compound Annual Rates of Growth					

Sources: Joint Economic Committee of Congress (1986, table 4, p. 80, and table 5, p. 81); Bergson (1978, appendix table 11, p. 236).

Method: Factors are combined, following Bergson, according to the Cobb-Douglas specification (see Bergson, 1978, pp. 159–60, fn. 8). The income elasticities (shares) inputed to capital (0.328 GNP, 0.317 industry) are provided in Bergson, (1978, appendix table 11). Labor is 0.672 GNP, and 0.683 industry. The capital weights applied measure net reproducible capital services, assuming that the CIA's series that is not defined grows proportionally with the net stock. Similar results are generated using Bergson's gross, enterprise capital stock weights (1978, appendix table 22, p. 245). Both the input and output data are taken from CIA sources. Total factor productivity growth for aggregate output and industry are computed with the formula

$$\dot{\theta} = \dot{y} - [\alpha\dot{k} + (1-\alpha)\dot{l}],$$

where $\dot{y}$, $\dot{k}$, and $\dot{l}$ are output, capital, and labor growth, and α and $(1-\alpha)$ are the output elasticities of these factors measured in 1955 rubles (GNP) and 1959 rubles (industry).

partly on official MBMW growth series indicate that Soviet machine-building factor productivity may have surpassed the U.S. level in 1975 (Tables A12 and A16 in statistical appendix, available upon request).[13]

> We must also reckon perhaps with a degree of concealed inflation owing to overevaluation of new domestic machinery, though the evidence on that is particularly opaque. It would, I think, be surprising if the resultant overstatement overall in annual producers' durables investment growth was more than one percentage point 1971–75 and three percentage points for 1976–80.
>
> In sum, that the TsSU and CIA data on real fixed investment are subject to concealed inflation has yet to be conclusively demonstrated. But we must, I think, reckon those measures to be very possibly so affected. Should they be, the rate of growth cited above 1971–75 might have

[13] In evaluating the possibility that Soviet MBMW productivity may exceed the American level, one must carefully consider the extent to which official data are distorted by hidden inflation. Bergson appears to believe that hidden inflation is not too severe in some civilian MBMW activities, but he has not clarified his views on concealed inflation in the military machine-building sector. As a consequence, the degree to which he might be prepared to accept the official MBMW growth series after 1975 is unclear.

The trend thereafter is less certain. Soviet MBMW productivity deteriorates using the CIA's output estimates, and improves according to official series. These statistics in their totality are difficult to interpret, but they do suggest that factor productivity in the Soviet machine-building sector was at, or above, both the aggregate and industrial levels in 1975 and that the margin may have widened thereafter.

As previously cautioned all comparative assessments of Soviet economic performance are clouded by the suspicious character of Soviet statistics. It should also be observed that the CIA's dollars-ruble ratios on which Bergson and the author rely may overstate the quality of Soviet equipment. Nonetheless, to the extent that Bergson's methodological conventions are used as a norm, it follows that Soviet machinery and industrial productivity in 1975 were much closer to the American level than the economywide average. Whatever deficiencies beset the socialist economic systems of the East generally, they appear to operate with reduced force in industry and industrial subsectors supporting priority Soviet activities.[14]

II. Sources of Inefficiency

During the course of his various analyses, Bergson identifies numerous causes that partly explain Soviet and/or socialist economic underefficiency. He places the most importance on Soviet technological shortcomings[15] and the least on intraenterprise technical inefficiency. (Compare Padma Desai, 1987; Yasushi Toda, 1986; and Robert Whitesell and H. Barreto, 1986.) The finding that the productivity structure of the Soviet economy parallels the duality long manifest in its output structure does not contradict this causal ranking, but it does require further elaboration. Apparently, the potency of some or all these causes varies across branches and sectors. Recent empirical work on intraenterprise efficiency in Soviet cotton refining (Rosefielde, Knox Lovell, Vyachaslav Danilin, and Ivan Materov, 1985) and the confectionary sector (Mikhail Afanas"ev and V. Sokov, 1985; Afanas"ev, 1987), insofar as it is representative, indicates that technical inefficiency in light industry may not be a major source of aggregate underproductivity. Counterpart studies in other sectors have not been undertaken, but it would hardly be surprising if technical efficiency in priority activities were not at least up to the cotton refining and confectionary industrial norm (Rosefielde and R. W. Pfouts, 1988; Richard Ericson, 1984). It may therefore be provisionally inferred that the superior productivity of the Soviet industrial sector generally, and the machine-building sector in particular, are attributable to the selective acquisition of advanced technologies and the priority allocation of key factor supplies. Priority sectors enjoy privileged access to scarce, superior inputs, have some influence on the production of intermediate input supplies including special order goods, and are less prone to factor supply disruptions. Research and development, like technology transfer efforts, are concentrated in these activities and are more effectively diffused because of the close working arrangements among foreign trade organizations, specialized procurement agencies, design bureaus and enterprises. The asymmetries in measured

> to be discounted by a fraction of a percentage point. For the tempo cited 1976–80, the corresponding discount might be, say, a full percentage point. [Bergson, 1987c, p. 420]

In his fn. 16 Bergson explains further that hidden inflation in the MBMW producers' durables sector is likely to be in line with rates on fixed investment because TsSU's producers' durables statistics do not appear to be biased by imported foreign durables prices. Readers wishing to explore this matter further are directed to the appendix tables cited, CIA (1988), and Rosefielde (1988a, 1988b).

[14] In appraising the performance of the Soviet machine-building sector it should be borne in mind that the U.S. Defense Intelligence Agency considers the technology of deployed Soviet weapons to be on a par with America's. See Joint Economic Committee of Congress, 1987, p. 88.

[15] Bergson, 1987a, pp. 355–56. Also see his discussion on prices, p. 345.

factor productivity manifested by the Soviets and other socialist economies thus may not be accidental. They may reflect enduring institutional priorities of diverse sorts that, other things equal, should be expected to persist, thwarting all efforts to prove on purely positive grounds that these economies are, or must eventually be, comprehensively underefficient.

III. Conclusion

Subject to the usual controversies regarding Soviet data and profound but intractable difficulties of measuring comparative demand satisfaction, this note supports Bergson's important finding that the Soviet economy and socialist economies of the East were statistically underproductive (underefficient) in 1975. However, it disconfirms two closely related subhypotheses. First, contrary to the spirit of Bergson's article, his own data demonstrate that the Soviets narrowed the productivity gap with the West in 1960–75 and that socialism as it has developed in the Soviet Union and Eastern Europe therefore need not always be measurably underproductive in all regards. Second, disaggregation revealed further that factor productivity in the Soviet economy exhibits a pronounced duality that may make it possible for the Soviets to excel in some aspects of industrial production, while lagging noticeably in others. These findings cast some doubt on the adequacy of the kinds of adjusted data on which Bergson relies, but insofar as they serve as a valid norm, it must be concluded that the socialist economies of the East are not unambiguously and comprehensively underproductive (underefficient). An impartial reading of the evidence thus indicates that Soviet economic performance evaluated sectorally and/or dynamically may not always be inferior to its rivals in the developed West, and might be improved by administrative, incentive, property right, and/or market reforms (Aleksandr Isaev, 1989). Although the prognosis for Soviet and Eastern European productivity (efficiency), given their poor performance since 1975 and the widespread dissatisfaction that has been expressed in the East with the quality of socialist life, is hardly encouraging, it is still premature to suppose that the aggregate productivity of these socialist economies will continue to underperform the capitalist economies of the developed West according to the measures Bergson prefers, or that the East cannot excel in some important industrial pursuits.

REFERENCES

Afanas"ev, M., "Metody otsenki effektivnosti proizvodstva dlia obosnovaniia planovykh reshenii," in S. Rosefielde et al., *Theory and Measurement of Economic Efficiency in the Soviet Union and the United States*, unpublished manuscript, 1989.

______ **and Sokov, V.,** "The Estimation of the Efficiency of Enterprise Activities," in Pekka Sutela, ed., *Proceedings of the 6th Finnish-Soviet Symposium in Economics*, Helsinki, 1985, 77–94.

Bergson, Abram, "Socialist Economics," in H. Ellis, ed., *A Survey of Contemporary Economics*, reprinted in Abram Bergson, *Essays in Normative Economics*, Cambridge, MA: Harvard University, Belknap Press, 1948, 193–206.

______, *Planning and Productivity Under Soviet Socialism*, Pittsburgh, PA: Carnegie Mellon University, 1968.

______, *Soviet Post-War Economic Development*, Stockholm, Sweden: Almqvist & Wiksell International, 1974.

______, *Productivity and the Social System —The USSR and the West*, Cambridge, MA: Harvard University Press, 1978.

______, "Technological Progress," in A. Bergson and H. Levine, eds., *The Soviet Economy: Toward the Year 2000*, London: Allen & Unwin, 1983, 34–78.

______, (1987a) "Comparative Productivity: The USSR, Eastern Europe, and the West," *American Economic Review*, June 1987, 77, 342–57.

______, (1987b) "Comparative Productivity: USSR, Eastern Europe and the West: Appendix Sources and Methods for Basic Data," unpublished manuscript, 1987.

______, (1987c) "On Soviet Real Investment Growth," *Soviet Studies*, July 1987, *39*, 406–24.

Birman, Igor, *Ekonomika nedostach*, New York: Chalidze Publishers, 1983.

Desai, Padma, *The Soviet Economy: Problems and Prospects*, Oxford: Basil Blackwell, 1987.

Edwards, Imogene, Hughes, Margaret and Noren, James, "U.S. and U.S.S.R.: Comparisons of GNP," *Soviet Economy in a Time of Change*, Vol. 1, Washington: Joint Economic Committee of Congress, 1979, 369–401.

Ericson, Richard, "The 'Second Economy' and Resource Allocation Under Central Planning," *Journal of Comparative Economics*, March 1984, *8*, 1–24.

Gallik, Dmitri, Kostinsky, Barry and Treml, Vladimir, *Input-Output Structure of the Soviet Economy in 1972*, Foreign Economic Report, No. 18, Washington: Bureau of the Census, Department of Commerce, 1983.

Isaev, Aleksandr, "Reforma i oboronnye otrasli," *Kommunist*, Mart 1989, *3*, 24–30.

Rosefielde, Steven, *The Transformation of the 1966 Soviet Input-Output Table from Producers to Adjusted Factor Cost Values*, GE75TMP-47, Washington, DC: General Electric TEMPO, 1975.

______, *East-West Trade and Postwar Economic Growth*, Washington, D.C.: Stanford Research Institute, 1976.

______, *False Science: Underestimating the Soviet Arms Buildup*, 2nd ed., New Brunswick, NJ: Transaction, 1987.

______, (1988a) "The Soviet Economy in Crisis: Birman's Cumulative Disequilibrium Hypothesis" *Soviet Studies*, April 1988, *40*, 222–44.

______, (1988b) "Soviet Hidden Hyper-Inflation: Conjecture and Knowledge," Paper presented at the AAASS meetings, Honolulu, November 20, 1988.

______, (1988c) *Tempting Fate: Economic Foundations of Soviet National Security Strategy*, unpublished manuscript, 1988.

______, (1988d) *Annotated Compendium of Soviet Economic Statistics 1950–85*, unpublished manuscript, 1988.

______, **Lovell, K., Danilin, S. and Materov, I.,** "Measuring and Improving Enterprise Efficiency in the Soviet Union," *Economica*, May 1985, *52*, 225–34.

______ **and Pfouts, R. W.,** "Economic Optimization and Technical Efficiency in Soviet Enterprises Jointly Regulated by Plans and Incentives," *European Economic Review*, July 1988, *32*, 1285–99.

Toda, Yasushi, "The Imperfections in the Factor Markets and the Loss of Production on Centrally Planned Industry: A General Equilibrium Calculation," Paper presented at the 18th American Association for the Advancement of Slavic Studies Meetings, New Orleans, November 20–23, 1986.

Whitesell, R., and Barreto H., "Estimation of Soviet Output Loss and Allocative Inefficiency: A Comparison of Estimates," unpublished manuscript, September 1986.

CIA, *USSR: Gross National Product Accounts, 1970, Research Aid*, DDB-1900-140-87(*U*), 1975.

______, *USSR: Measures of Economic Growth and Development, 1950–80*, Washington: Joint Economic Committee of Congress, 1982.

______, *A Comparison of Soviet and U.S. Gross National Products 1960–83 (U)*, SOV-84-10114, (S), 1984.

______, *The Impact of Gorbachev's Policies on Soviet Economic Statistics*, SOV88-10048 (U), 1988.

DIA, *Gorbachev's Modernization Program: A Status Report*, DDB-1900-140-87, August 1987.

Joint Economic Committee of Congress, *Allocation of Resources in the Soviet Union and China—1985*, Washington: March 19, 1986.

______, *Allocation of Resources in the Soviet Union and China—1986*, Washington: March 19 and August 3, 1987.

U.S. Department of Commerce, *Fixed Reproducible Tangible Wealth in the United States, 1925–85*, Washington: Bureau of Economic Analysis, June 1987.

Comparative Productivity: Reply

By ABRAM BERGSON*

By reprocessing comparative productivity data for 1975 in Bergson (1987) to conform methodologically to related data for 1960 in Bergson (1972, reprinted 1978), Steven Rosefielde usefully provides a basis to gauge relative productivity shifts over time. He considers the advance in Soviet status thus manifest, however, as "remarkable." It is also held to represent one of several ways in which my principal finding, that Soviet factor productivity is depressed by Western standards, is misleadingly incomplete. Rosefielde apparently accepts that finding but concludes that "contrary to the spirit" of Bergson (1987), "the Soviets narrowed the productivity gap with the West in 1960–1975."

Perhaps they did, but if so, according to his compilation, the narrowing was hardly dramatic. At 41 (USA = 100), Soviet factor productivity in 1960 is 60.5 percent of the average for the five Western countries considered. At 55, the corresponding 1975 ratio is 62.8 percent.

As recent writings of William Baumol and others show, the quantitative relation between "development stage" and productivity growth is a complex matter. According to a separate inquiry (Bergson, 1974, reprinted 1978), however, factor productivity growth during 1955–70 varied markedly among the same five Western countries, plus Japan. The variation takes the form of a statistically significant inverse relation to development stage, such as often hypothesized. The calculations also indicate, though with somewhat less statistical assurance, what perhaps might have been inferred from Rosefielde's compilation, that considering development stage, the USSR here too is relatively depressed.

In comparing factor productivity levels in the USSR and the West in Bergson (1972), I compiled data not only for all "material sectors" together but for "industry" alone. The latter calculation, however, proved especially laborious, and in Bergson (1987), with a decidedly larger sample of countries to deal with (including, in addition to the six studied for 1960, five others: Japan, Spain, Poland, Hungary, and Yugoslavia), I focused on the more comprehensive sphere alone. In view of the importance that Rosefielde attaches to this deficiency, perhaps he himself will feel impelled to repair it.

Meantime, though, he properly states my finding that, regarding factor productivity, the USSR in 1960 compares more favorably with the West in industry than in material sectors generally. As he seems to acknowledge, though, with appropriate normalization, the USSR probably underperforms the West in one sphere as well as in the other (Bergson, 1972).

Rosefielde must nevertheless be right to infer that factor productivity in industry in the USSR is higher than in the economy more generally. Whether that margin of superiority has been increasing, as he argues, is doubtful. In view of inevitable data limitations, the edge in favor of industry shown by his calculations seems rather narrow. Also, as he fails to consider, his measures of productivity growth in the economy generally cover services. For purposes of productivity measurement, Bergson (1972, 1987) rightly focuses on material sectors. There, factor productivity growth has been distinctly greater than in the economy generally and probably has been fully comparable to that in industry (Bergson, 1974, 1983). Contrary to Rosefielde's further speculation, industrial factor productivity in the USSR may have retrogressed relative to that in the West during 1955–70 (Bergson, 1974).

Rosefielde's computations for MBMW are novel and interesting. He is probably right

*Department of Economics, Harvard University, Cambridge, MA 02138.

that Soviet productivity compares more favorably with that of the USA there than in industry as a whole. He might have stressed more than he does (fn. 13), though, that MBMW includes military machine building. For fairly obvious reasons, comparative data on output in this sphere must be subject to decidedly more than the usual margin of error attaching to such statistics.

Rosefielde alludes more than once to the dubious nature of "Soviet statistics" underlying my comparative productivity computations. As the nonspecialists may not grasp, he is being cautioned thus not primarily about Soviet official statistical data, such as have rightly come to be discredited among Soviet as well as Western scholars (and that Rosefielde himself audaciously uses in the second of his two computations on MBMW). Rather, Rosefielde is referring in that odd way to widely accepted Western recalculations of such Soviet data. Substantial and carefully documented as such recalculations are, however, they too have their limitations. In cited studies, I have sought systematically to address such limitations. My findings, I believe, are quite robust to resultant data errors.

Rosefielde speculates finally about the conditioning factors underlying Soviet underperformance regarding factor productivity. My discussions of this interesting question in a number of different writings admittedly call for collation and review, but that is a task for another occasion. Flawed as Rosefielde's critique is, it effectively underscores the complexity of comparative economic performance appraisal.

REFERENCES

Bergson, Abram, "Productivity Under Two Systems: USSR and the West," in Abram Bergson, Jan Tinbergen, Fritz Machlup, and Oskar Morgenstern, eds., *Optimal Social Welfare and Productivity*, New York: New York University Press, 1972, 53–104.

______, *Soviet Post-War Economic Development*, Stockholm: Almquist & Wicksell, 1974.

______, *Productivity and the Social System—The USSR and the West*, Cambridge, MA: Harvard University Press, 1978.

______, "Technological Progress," in Abram Bergson and Herbert S. Levine, eds., *The Soviet Economy: Toward the Year 2000*, London: Allen & Unwin, 1983.

______, "Comparative Productivity: The USSR, Eastern Europe, and the West," *American Economic Review*, June 1987, 77, 342–57.

On the Basing-Point System: Comment

By David D. Haddock*

In "On the Basing-Point System," Bruce Benson, Melvin Greenhut, and George Norman (1990)[1] properly correct a misunderstanding by Jacques Thisse and Xavier Vives (1988)[2] concerning my 1982 paper on basing-point prices. But BGN then take exception with several of the conclusions I had reached, or that they thought I had reached. Although I have insufficient space to deal with all the points they raise, I will address the most important ones.

In 1981 most economists believed that a model of profit-maximizing basing-point pricing would have to assume collusion. My model, as BGN seem to agree, showed that it need not; atomistic competition at one production site accompanied by local monopolies elsewhere may lead to basing-point pricing and freight absorption without collusion.[3] Conversely, TV showed that the practice as observed empirically will not arise if all production sites contain noncooperative local monopolies. BGN establish yet another important theoretical result—noncooperative oligopoly at one site could generate the f.o.b. prices (or f.o.b. plus transport) that, in my model, led local monopolies elsewhere to adopt freight-absorbing basing-point prices. The three models would seem to be complements.

I. A Paradox

But BGN claim that their formalization shows that noncooperative freight-absorbing basing-point prices require "highly restrictive conditions" (p. 584), implying to them that such systems would be rare. Hence, BGN contend that, though the logic of such a model is internally consistent, rivalrous freight-absorbing basing-point systems are rare or nonexistent in practice; actual instances must arise from conscious parallelism or collusion (p. 587). And even if such prices reflected rivalrous behavior, the implied spatial price discrimination might be welfare reducing.

BGN's contention creates a paradox: (*1*) BGN claim that basing-point pricing with freight-absorption is not rivalrous in practice; (*2*) economists have long suspected that a strong cartel will not use freight-absorbing basing-point prices; (*3*) hence, if (*1*) and (*2*) are correct, the practice must arise in some sort of *weak* cartel; but (*4*) neither BGN nor any earlier writer has elucidated a widely applicable and satisfactorily rigorous model that implies that basing-point pricing will comprise profit-maximizing behavior by a weak cartel.

It seems peculiar to try to settle an essentially empirical issue through pure theory, particularly since BGN offer no comparably rigorous model that yields basing-point prices as a collusive practice.[4] Lacking the

*Northwestern University School of Law, Chicago, IL 60611-3069. I am grateful to Ian Ayres, Marty DeBoer, Mark Grady, Fred McChesney, and Menahem Spiegel for helpful comments.

[1] BGN henceforth.

[2] TV henceforth.

[3] On the other hand, the situation envisioned may lead to some alternative pricing pattern. Under some circumstances the profit-maximizing monopoly price for a remote firm will be less than a price based on those of atomistic firms elsewhere. In that event, the remote firm will charge a monopoly price. If it is too costly for the remote firm to price discriminate spatially in a more refined manner, the remote firm will become a base in a multiple basing-point system. See Section III, part B of my 1982 article.

[4] In the latter regard BGN resemble TV, who remark that a single base system is "not a stable configuration since it is not an equilibrium of our two-stage game. This suggests the hypothesis that [basing-point pricing] cannot be explained in the context of a noncooperative model...and suggests that theoretical explanations of [basing-point pricing] should consider its role as a coordinating and collusive device..." (pp. 130–31). That suggestion, however, is reached by default, because TV have no collusive model yielding basing-point pricing either. As BGN properly note, TV were simply not working with a model capable of

latter model, it would seem impossible to guess, even on theoretical intuition, whether or not collusive basing-point prices require less restrictive conditions than do rivalrous ones.

Nevertheless, for the sake of argument, assume BGN are correct. What would that imply? Freight-absorbing basing-point pricing seems to be an unusual system. Few industries ever use it, and then only transitorily.[5] If the circumstances that implied such a rare practice were not restrictive, then one might indeed suspect collusion. Hence, I find no cause in BGN to alter my prior belief that basing-point prices per se, even those involving freight-absorption, provide no *prima facie* evidence about collusion one way or the other. If most firms are not (are) colluders, then most firms using basing-point prices are (are not) probably exhibiting rivalrous behavior.

Clearly, that does not mean that a firm utilizing basing-point prices can never be guilty of collusive behavior, just as f.o.b. pricing (or any third pricing structure) does not prove rivalrous practice.[6] But as I indicated in 1982 (p. 293), those alleging collusion should be required to produce other evidence, because basing-point prices themselves seem useless as indicators.[7]

II. Refuting the Presumption of Collusion

A. *Limit-Entry Location*

BGN base their belief that cartels might adopt basing-point prices on the hypothesis of pure profit with freedom of entry, as advanced by Nicholas Kaldor (1935) and developed by Curtis Eaton and Richard Lipsey (1978).[8] The Kaldor-Eaton-Lipsey model resembles a spatial version of limit-entry pricing. Plants locate closely enough together to enable the existing ones to earn some profits, though rarely short-run profit-maximizing ones, but there is insufficient demand to cover average costs in the event of entry.[9] BGN argue that "more

generating basing-point pricing, but that does not mean one cannot exist—my rivalrous one, for example, or BGN's oligopolized variant.

Curiously, immediately following the quoted passage, TV very briefly outline in rudimentary form alterations in their model that would yield conclusions broadly consistent with my own.

[5] Except, perhaps, for some European examples. But there the pricing structure is established and enforced through governmentally established bureaucracies.

[6] In fact, I noted in 1982 that "basing-point prices may arise in a competitive industry *at least as readily as in a poorly cartelized one*" (p. 293, emphasis added). Only the competitive model was elucidated; I had nothing to add to the existing literature regarding the poorly cartelized possibility, as formulated by George Stigler (1949) and John McGee (1954). Subsequent publication of Dennis Carlton's 1983 article, however, has dramatically weakened my confidence in the prior theory of poorly cartelized basing points.

[7] One collusive organization that might opt for basing-point prices is one that is strong locally but has geographically limited influence—that is, one that cannot achieve agreement with, detect chiseling by, and/or punish violations of the cartel agreement by at least some firms that are at a distance. Functionally, a locally strong cartel resembles a spatial monopolist, and so resembles the remote firm of my 1982 article. A simple extension of my 1982 model shows that under the appropriate conditions maximizing cartel profits will require basing cartel members' prices on those of nonmembers elsewhere, and then absorbing freight. See Section VI, Part B of my 1982 article for a discussion of behavior that may have reflected the use of such weakly collusive basing-point pricing. Also see *Federal Trade Commission* v. *National Lead Company et al.* In 1933, while drafting a "code of fair competition" for themselves under the National Industrial Recovery Act, several lead producers adopted a spatial price schedule that resembled that of the industry's largest firm.

Beginning with a similar insight, Thomas Gilligan (1990) has independently undertaken a detailed theoretical and empirical investigation of the plywood industry.

Note carefully, however, that the requisite conditions for this form of weakly collusive basing-point pricing closely resemble the ones that lead remote firms to absorb freight in conjunction with basing-point pricing in a rivalrous market. Consequently, from a purely theoretical standpoint, those collusive requirements are no easier to satisfy than are those facing my rivalrous model. Whether or not the use of basing-point prices reflects collusion, then, remains an empirical issue that requires independent evidence of collusion for its resolution.

[8] BGN, p. 586.

[9] Though resembling limit-entry price theories, the Kaldor-Eaton-Lipsey hypothesis is on firmer ground when firms have site-specific long-lived investments as credible precommitments to "stand their ground" in the face of entry-induced losses, a point alluded to by Eaton and Lipsey, and developed by Dennis Capozza and Robert Van Order (1980).

profitable systems may well induce competitive entry. ...[I]f base-point pricing...protects existing rents, it may actually be the most desirable arrangement...[because]...a base-point system is much less costly to police and enforce..." (p. 588).

Kaldor, Eaton, and Lipsey have not gone unchallenged. The erosion of abnormal gains when the rights to them are ill-defined is well understood for other activities.[10] Ronald Johnson and Allen Parkman (1983) argued that the same "tragedy of the commons" applies to Kaldor-Eaton-Lipsey. Expected future profits for incumbents induce premature investment. Profits aggregated over the investment's life will be normal, though the plant incurs losses during some periods and reaps profits during others.

So entry cannot be deterred by Kaldor-Eaton-Lipsey techniques unless an industry is stagnant or declining. Indeed, Johnson and Parkman claim to refute the empirical implications of the hypothesis for the cement industry. If basing-point prices will not deter entry, one would expect a cartel—whether weak or strong—simply to maximize short-run profits, but that will usually require a pricing structure other than one using basing points.

If property rights to above-normal gains *are* well defined, as through ownership of essential inputs or processes, or through regulation, the rights owners can control entry directly, if that is desirable. The owners can simply refuse to sell the input, license the process, or authorize the entry, and thus have no incentive to adopt basing points to exclude potential competitors.

B. *A Cartel's Ability to Detect Chiseling*

Not only do basing-point prices not maximize cartel profits, entry possibilities being neglected, they actually enhance opportunities to chisel. Contrary to the supposition of many observers, BGN included, basing-point prices interfere with cartel discipline. George Stigler (1949) is the most cited author on collusive basing points. But Carlton argues that Stigler's own "A Theory of Oligopoly" (1964) is inconsistent with Stigler's earlier theory of delivered prices. A cartel must first detect chiseling to control it. If prices are unobservable, Stigler's oligopoly theory shows that a second-best means to detect chiseling is to observe sales patterns. Changes that cannot plausibly reflect stoachastic factors will signal chiseling.

Basing points make price irrelevant for buyers' choices among sellers, the sellers seem too homogeneous to buyers, so chiseling becomes more difficult to infer from a given shift in buying patterns. But many alternative pricing systems yield similar indeterminacy only along market boundaries between adjoining production sites. Chiseling elsewhere will create damning evidence. Even if all colluders at a single site quote identical prices to every customer because no other way to divide the *local* market has proven workable, they will retain market boundaries with neighboring cartel members. Market division is a better cartel strategy than is product homogenization. To argue the contrary implies that firms prefer to be oligopolists rather than monopolists, or that they prefer to be n-firm oligopolists rather than m-firm oligopolists, where $n > m$.

If, in contrast, prices are observable, basing points are not inferior to other systems for detecting chiseling. But neither are they superior. Indeed, no pricing system is disadvantaged for that purpose. But some alternative system will typically yield higher cartel profits than will basing points. So neither strong nor weak cartels would often use basing points, providing the cartels are industrywide.

C. *Procollusive Legal Attacks on Basing-Point Pricing*

Alternatively, suppose a cartel exists that does not include all the firms in the industry, or one that has difficulty policing its members' prices. Contrary to BGN's assertion, Carlton shows that the cartel would

[10] H. Scott Gordon (1954) was an early expositor of the "property rights paradigm." See Terry Anderson and P. J. Hill (1990), and Richard Cornes, Charles Mason, and Todd Sandler (1986) for updated bibliographies.

like to divide market areas among cartel members, not adopt basing-point prices. Assume that the cartel tries to institute a pricing scheme that divides markets.[11] Since the cartel's price quotations exceed marginal cost (if the cartel merits concern), a firm at one site can profitably invade all or part of another site's "natural market" by matching (or slightly undercutting) cartel prices there. The invasion may reflect chiseling by a cartel member, or simply opportunism by a nonmember. The chiseling/nonmember firm would, in effect, be creating a multibase basing-point system if the cartel had adopted f.o.b. prices.

But in that event the cartel itself may instigate a challenge to the multibase system by encouraging some buyer, the Federal Trade Commission (FTC), or the Justice Department to sue the chiseling/nonmember firm for "anticompetitive" behavior. If the suit is successful, the chiseling/nonmember firm will be forced to limit its anticartel behavior to its own natural market area. In other words, antitrust suits against basing-point prices can be *pro*collusive actions.

III. Welfare Effects of Rivalrous Basing-Point Pricing

Even if freight-absorbing basing-point systems appear in rivalrous industries, BGN worry that consumers gain nothing. Perhaps the enhanced revenues become rents, locational or otherwise, or (I might add) are dissipated in needless overhead. I agree with the spirit of that suggestion. Whether the discrimination increases welfare ultimately is another empirical matter, and another one that will not be resolved here. Nevertheless, there are several additional points that must be underlined.

First, an "Austrian" economist might foresee economic advantage whether or not increased consumer surplus results. If all potential entrants are inferior to a firm that absorbs freight under a basing-point schedule, the firm may indeed earn rents. That the firm would necessarily survive if forced to adopt a less preferred pricing schedule does not follow, however. And producer surpluses count, too.

And even neoclassical models offer potential sources of consumer surplus that ought not to be ignored. Arthur R. Burns (1936), no apologist for basing-point prices, nevertheless often observed *near*-basing-point prices, not *precise* ones;[12] to capture the bulk of sales in its "natural market," the remote firm in my model must be superior to distant firms in the eyes of local buyers. But discounts provide buyers a benefit. BGN might think it a paltry benefit, but perhaps it is the greatest one possible if the firm is a high cost producer (transport cost aside).[13]

[11] An implication of an article by Arthur Smithies (1941) is that a cartel would often opt for some geographically discriminatory schedule, though not basing-point prices, rather than f.o.b. prices. That distinction is immaterial to Carlton's discussion; many price schedules other than f.o.b. would divide markets.

[12] This observation was briefly discussed in my 1982 article at fn. 8.

"Near-basing-point price" and "precise-basing-point price" are not terms coined by Burns, but accurately denote his findings. Burns was specifically interested in steel pricing, but ample evidence indicates near-basing-point prices have been the rule rather than the exception. For example, in *Federal Trade Commission* v. *Staley Mfg. Co.* it was noted that for glucose sales the "[r]espondents, to gain access to the markets...made their sales 'by first quoting the same prices as were quoted by competitors and then making whatever reduction in price...was necessary to obtain business'" (Supreme Court of the United States, quoting, in part, evidence gathered by the FTC, 324 U.S. 746, 751). Similarly, Greenhut (1987) remarks that for plywood "the new entrants in the South established a delivered pricing formula which assured the capture of the southern market. They accomplished this by quoting a Portland Base price at a slight differential 'below' the f.o.b. mill price on Doulgas Fir plus West Coast freight to southern destinations" (p. 202).

[13] One is reminded of the outrage that greeted Robert Fogel (1964) when he discovered that the U.S. rail system had yielded a social saving of "only" a percent or two of GNP. Fogel's response, of course, was that one is hard-pressed to find *any* alteration in economic structure that results in really staggering benefits to the economy. If one dismisses seriatim every "small" increment in economic welfare, there is little left to count.

Further, the firm's effort to offer only a modest discount will sometimes fail if the firm must rely on buyers' representations to estimate competing prices, for buyers have a clear incentive to misinform sellers. Hence, some buyers will obtain discounts that are more than modest.[14]

Finally, suppose for the sake of argument that rivalrous freight-absorbing basing-point prices yield gains in neither consumer nor producer surpluses. Even a zero gain is preferable to a loss. There is little question that either regulatory actions or private litigation absorb a substantial amount of real resources and, worse yet, can often be turned opportunistically to nefarious, socially injurious purposes, as argued by Stigler (1971) and his followers. Antitrust actions should be reserved to remedy actual injuries, not to attach parties that (some may think) have rendered merely modest benefits. It is unwise to insist that courts take a more active role in fine-tuning the operations of complex commercial enterprises. Justices have JDs, not MBAs; in a healthy society courts are asked to render justice, not to administer firms.

The welfare impact of freight-absorbing basing-point prices will not be discovered solely by theoretical means. Hard empirical work is called for at this juncture.

[14] Recognizing buyers' incentives to mislead, a firm will sometimes refuse to undercut a claimed offer. But if the claim is genuine, a failure to underbid will cause the order to be lost to a more distant rival; there will be both Type I and Type II errors, which is merely an implication of Stigler (1961). A seller's inability to perfectly separate genuine buyers' reports from opportunistically misleading ones can account for at least some intermingled sales in freight-absorbing industries, as reported by Fritz Machlup (1949); just as it accounts for intermingled sales across natural market boundaries in f.o.b. industries. Additional intermingling, it was argued by Stigler (1949), arises from lags between orders and deliveries when prices vary.

There are other, even more subtle, benefits to buyers that may enable the remote firm to dominate its local market. For example, the firm may be able to offer quicker delivery or more personalized service or, though seeming to quote precise basing-point prices, be less insistent on prompt payment. These are but a few possibilities.

REFERENCES

Anderson, Terry L. and Hill, P. J., "The Race for Property Rights," *Journal of Law and Economics*, April 1990, *33*, 177–97.

Benson, Bruce L., Greenhut, Melvin L. and Norman, George, "On the Basing-Point System," *American Economic Review*, June 1990, *80*, 584–88.

Burns, Arthur Robert, *The Decline of Competition*, New York: McGraw-Hill, 1936.

Capozza, Dennis R. and Van Order, Robert, "Unique Equilibria, Pure Profits, and Efficiency in Location Models," *American Economic Review*, December 1980, *70*, 1046–53.

Carlton, Dennis W., "A Reexamination of Delivered Pricing Systems," *Journal of Law and Economics*, April 1983, *26*, 51–70.

Cornes, Richard, Mason, Charles F. and Sandler, Todd, "The Commons and the Optimal Number of Firms," *Quarterly Journal of Economics*, August 1986, *101*, 641–46.

Eaton, Curtis B. and Lipsey, Richard G., "Freedom of Entry and the Existence of Pure Profit," *Economic Journal*, September 1978, *88*, 455–69.

Fogel, Robert William, *Railroads and American Economic Growth*, Baltimore, MD: Johns Hopkins University Press, 1964.

Gilligan, Thomas W., "Imperfect Competition and Basing-Point Pricing: An Application to the Softwood Plywood Industry," unpublished manuscript, Hoover Institution, March 1990.

Gordon, H. Scott, "The Economic Theory of a Common-Property Resource: The Fishery," *Journal of Political Economy*, April 1954, *62*, 124–42.

Greenhut, M. L., "Basing Point System," in John Eatwell, Murray Milgate, and Peter Newman, eds., *The New Palgrave: A Dictionary of Economics*, Vol. 1, New York: Stockwell Press, 1987.

Haddock, David D., "Basing-Point Pricing: Competitive vs. Collusive Theories," *American Economic Review*, June 1982, *72*, 289–306.

Johnson, Ronald N. and Parkman, Allen, "Spatial Monopoly, Non-Zero Profits and Entry Deterrence: The Case of Cement," *Review of Economics and Statistics*, August 1983, *65*, 431–39.

Kaldor, Nicholas, "Market Imperfection and Excess Capacity," *Economica*, February 1935, *2*, 35–50.

Machlup, Fritz, *The Basing Point System*, Philadelphia: Blakiston, 1949.

McGee, John S., "Cross Hauling—A Sympton of Incomplete Collusion Under Basing-Point Systems," *Southern Economic Journal*, April 1954, *20*, 369–79.

Smithies, Arthur, "Monopolistic Price Policy in a Spatial Market," *Econometrica*, January 1941, *9*, 63–73.

Stigler, George J., "The Economics of Information," *Journal of Political Economy*, June 1961, *69*, 213–25; reprinted in George J. Stigler, *The Organization of Industry*, Homewood, IL: Richard D. Irwin, 1968, 171–90.

______, "A Theory of Delivered Price Systems," *American Economic Review*, December 1949, *39*, 1143–59; reprinted in George J. Stigler, *The Organization of Industry*, Homewood, IL: Richard D. Irwin, 1968, 147–64.

______, "The Theory of Economic Regulation," *Bell Journal of Economics and Management Science*, Spring 1971, *2*, 3–21.

______, "A Theory of Oligopoly," *Journal of Political Economy*, February 1964, *72*, 44–61; reprinted in George J. Stigler, *The Organization of Industry*, Homewood, IL: Richard D. Irwin, 1968, 39–63.

Thisse, Jacques-Francois and Vives, Xavier, "On the Strategic Choice of Spatial Price Policy," *American Economic Review*, March 1988, *78*, 122–37.

Federal Trade Commission v. National Lead Company et al., 352 *US* 419–31, *decided February* 25, 1957.

______ **v. Staley Mfg. Co.,** 324 U.S. 746–760, decided April 23, 1945.

On the Basing-Point System: Reply

By Bruce L. Benson, Melvin L. Greenhut, and George Norman*

David D. Haddock's (1990) contention is really that the Jacques F. Thisse and Xavier Vives (1988) and the Bruce L. Benson, Melvin L. Greenhut, and George Norman (1990) papers do not provide a collusive model that yields base-point pricing (BPP). He apparently believes his paper (1982) established competitive BPP and suggests (in 1990) that the collusive model and hard empirical work alone remain. However, no collusive model is required since it has already been well demonstrated in the literature on plant location and spatial price theory that firms at feasible distance locations from a production center would accept BPP only because they fear that to do otherwise would subject them to retaliatory actions by the larger, more powerful firms at the production center. Moreover, the paper by Thisse-Vives (1988, henceforth TV) and Benson-Greenhut-Norman (1990, henceforth BGN) indicated sufficiently that other systems would arise than Haddock's competitive BPP, especially since the restrictive conditions necessary for the latter militate against its use. Haddock, in fact, appears to accept these conditions in his references to Arthur R. Burns' proposals (1936) of *near*-base-point pricing (NBPP). Hard empirical analysis would indeed be in order in determining the extent of NBPP, not competitive BPP.

This paper responds to the above noted issues in three main sections. Section I briefly provides some historical data on the collusive uses of BPP for any future researcher who seeks empirically to uncover a true competitive BPP system. Most importantly from a theoretical standpoint, this section also explains why cartels will employ BPP notwithstanding the obvious and well-known thesis that other more profitable spatial pricing forms exist, *ceteris paribus*. Section II of this paper then distinguishes between profits and locational rents, in the process indicating in partial contradiction to Haddock (1990) that BPP does generate such rents. Though we have already contended that existing theory has demonstrated the use of different competitive forms of spatial pricing *and* locations under profit-maximizing conditions than competitive BPP, Section III of this paper provides specific basis for this contention as well as references. That same section discusses NBPP. To save space, we shall let the analysis on the subject of the TV (1988) and BGN (1990) papers stand by themselves, without repeat demonstrations or explanations of their results.

I. What Is an Effective Cartel?

Haddock's references to weak or strong cartels and apparent suggestion (1990) that BPP would be used by weak cartels oversimplifies the subject. A cartel arises if it is able to cover the cost of organizing, and then survives if it establishes a sufficiently strong monitoring and enforcement (policing) system that is capable of limiting competition (for example, price competition, entry) to an acceptable level. Cartel costs vary with the characteristics of an industry (number of firms, geographic distribution of firms *and* consumers, etc.). When firms and consumers are geographically dispersed, organizational and policing costs tend to be high and the ability of the cartel to limit competition is lessened.

We echo many others when we suggest that BPP reduces policing costs for a geographically dispersed group. Haddock is not

*Department of Economics, Florida State University, Tallahassee, FL 32306-2045; Department of Economics, Texas A&M University, College Station, TX 77843-4228, and adjunct, University of Oklahoma, Norman, OK 73091; and Department of Economics, The University of Leicester, Leicester, England LE1 7 RH. The authors wish to thank Tom Saving for a helpful suggestion.

convinced by this argument. He contends that cheating will be detected more readily under f.o.b. pricing because prices (both f.o.b. and basing-point) are not easily observed. He proposes that effective monitoring must therefore involve observations of relative sales. He suggests that chiseling at a distance from a boundary "will create damning evidence" under f.o.b. pricing (1990, p. 959). We would propose, however, that prices are readily observed under many real basing-point pricing arrangements. Is Haddock suggesting that chiseling would not have been easily detected on the part of one or more of the eleven differentially located firms that offered cement to the U.S. government at the identical delivered prices of $3.286854 (*Aetna v. FTC*, 1946)? Or what of the ten sealed bids priced at $253,633.80 for reinforcing bars (*New York Times*, February 20, 1939), or the 59 steel pipe bids to the U.S. Navy Department, each for $6,001.83 (*Annual Report of the Attorney General*, 1937, pp. 37–38)? Or, consider the eight geographically dispersed cement companies that submitted bids to the Illinois Department of Highways for deliveries at 102 sites in 102 Illinois counties, every bid for every delivery being *identical* in price (*Congressional Record*, May 31, 1959, p. 7961).

The reason that basing-point prices are relatively easier to monitor is that they are determined by a simple formula that is known by every cartel member. For instance, between January 1982 and September 1983 there were three regional bases for the pricing of cement in England, and delivery increments were applied to every *5 miles of road*. These increments reflected haulage rate charges computed from a particular base point, *not the point of origin of the shipper*; these rates varied over time from 16p to 56p per ton per 5 miles of road. Prices certainly may not be readily observable under an f.o.b. system, but cooperative BPP makes them relatively more observable, and therefore reduces monitoring costs.

The fundamental feature of BPP that Haddock overlooks is that the main conspirators are typically localized at (near) the base point(s), and other spatially dispersed firms feel *obliged* to adhere completely to the system. A chiseler must be punished if caught, or there will be no effective deterrent to cheating. When chiseling under some organized type of pricing system other than BPP affects only a single firm located at a distant site, the incentives and ability of the cartel to retaliate and punish the transgressor are quite weak. This is particularly the case if the affected firm is small and not part of the main group of powerful conspirators. When chiseling impacts all cartel members, including the largest and most powerful, as would be the case under BPP, the potential for *and* likelihood of retaliation is much greater. This could conceivably induce, in some countries, the *pro*collusive type of antitrust action mentioned by Haddock; however, in a country such as the United States of the late twentieth century, any BPP cartel would be likely to pursue more subtle means of punishing a chiseler.

We contend that even if monitoring costs are no lower under a basing-point system than they are under some undefined, organized f.o.b. type of arrangement, *total* BPP cartel costs (which include enforcement as well as monitoring costs) will tend to be less. The imposition of credible punishment is an important part and cost of effective cartelization. For example, Clair Wilcox (1960) mentioned 49,000 pages of testimony along with 50,000 pages of exhibits on cement industry prices that were presented to the Federal Trade Commission. After also discussing the equally detailed documents on record concerning the steel industry, he went on to review penalties. Those imposed by the respective cartels on member firms that failed to adhere to the system were strikingly severe (pp. 280–81). He concluded that "If basing point pricing were a spontaneous outgrowth of natural causes, as some economists have argued, it would scarcely have been necessary to go to such lengths to ensure that its requirements be observed" (1960, p. 281).

Haddock is also incorrect when he suggests that sellers under BPP seem relatively more homogeneous to buyers than a cartel would wish (1990, p. 959). The fact of homogeneous output is exactly one reason why the system is desired by the large firms located at the base point. The steel, cement, and plywood firms that located at base points wanted *and took advantage of homogeneity* via a spatial delivered price that allowed them to maintain their markets over substantial distances. Haddock ignores an important aspect of many of these industries when he contends that a cartel would prefer geographic market division. As Wilcox (1960, p. 280) observed, changes in the geographic pattern of demand for steel (and presumably other construction material) took place rapidly in the United States. Thus, for these industries, producers' locations are fixed while consumers' are not.[1] When subject to geographic market division and f.o.b. pricing, firms would have to build new facilities to follow demand. This is not required as often under BPP since the large powerful firms under BPP can readily sell in new distant markets.[2]

II. Profits or Rents?

Haddock implies that cartelization must fulfill its objectives (for example, joint profit maximization, entry deterrence), or the cartel is ineffective. He refers to Ronald N. Johnson and Allen M. Parkman's (1983) demonstration that the cement industry did not earn supranormal profits, proposing this as evidence that any entry deterring efforts by the firms in the industry had to be ineffective. It would then seem to follow that any noncooperative, noncontrolled system could also be ineffective with respect to entry deterrence.[3] However, long-run rents in a true BPP system can appear in different form than higher net profits, while also resulting from entry deterring efforts. In particular, competition for prime locations would bid up the value of these sites. These locational rents would be capitalized as part of land values, rather than appearing regularly on P&L statements as high profits. More fundamentally, cement industry profits and individual firm profits are very different matters. This is especially the case because the distant small firm locates differently than a strictly competitive f.o.b. firm while, at the same instant, earning less under BPP than it otherwise would net (Melvin L. Greenhut, 1956). Finally, note that for risk-averse individuals, collusion that reduces the behavioral uncertainty inherent in noncooperative oligopoly could easily make the affected firms better off, even if nominal profits do not increase in any measurable way.

III. Conclusions: What in Fact Is Haddock's Noncooperative Spatial Pricing System?

Haddock's conclusions seem to agree with BGN that a true basing-point system would *not* exist under competitive conditions. Only its reflection would arise, with distant firms offering a modest discount.[4] He thus emphasizes what Burns (1936) called "near-base-point pricing," NBPP. If this is what Haddock means by competitive BPP, then

[1] BGN assume immobile consumers as well, but demonstrate that even in this case noncooperative pricing is very unlikely and market segmentation is the competitive result. In contrast, Haddock implies that market segmentation is strictly a collusive outcome.

[2] This is another consideration that would reinforce the TV (1988) conclusion that BPP becomes a relevant collusive practice.

[3] Haddock apparently accepts George J. Stigler's (1971) theory of economic regulation since he suggests that regulation and licensing may be a source of entry limits. According to this view, most regulatory actions provide cartel-like benefits for the regulated firms rather than benefiting consumers. Stigler's theory is based on the public choice paradigm in which bureaucrats and other public officials are driven by self-interest rather than public interest motives. Observe in this regard that the only examples of BPP which Haddock now appears to consider to be valid come from Europe "where the pricing structure is established and enforced through governmentally established bureaucracies" (1990, fn. 5). While we do not agree with Haddock that these are the only prime examples (for example, the federal milk order system in the United States has many basing-point characteristics), the fact that the most obvious examples arise in governmentally regulated settings reinforces the long-standing view that nongovernmentally imposed BPP is associated with cooperative rather than competitive behavior.

[4] A modest "discount" system could indeed prevail, and in a crude sense it often does in retailing where an uptown (suburban) department store *and* the uptown branches of downtown stores charge higher prices than the downtown price. Quite significantly, the differences in price correspond roughly to the time-distance "cost-saving" of suburban residents who shop at the stores located nearest them rather than having to go downtown.

there really is little disagreement between us except semantical. After all, competitive price discrimination, which just undercuts the base-point schedule, that is, NBPP, is precisely the noncooperative price equilibrium that arises in the BGN extension (1990) of the TV (1988) paper.

Competition in spatial markets, where both immobile buyers and sellers are spatially dispersed and transportation costs are significant, as modeled by Haddock (1982) and BGN (1990), leads naturally to a segmented market structure under which spatially separated firms, acting *independently*, would increase their profits by setting prices that undercut the distant rivals, *ceteris paribus*. It has, indeed, been well established in the literature on spatial price theory that freight-absorbing discriminatory pricing over a geographic space, as depicted by TV (1988) and BGN (1990), is the *natural pricing form* for noncooperative firms.[5] Certainly, demand elasticities can be expected on a priori grounds to differ at each buying site within a submarket (Edgar M. Hoover, 1936–37; Arthur F. Smithies, 1941; Greenhut, 1956). Note further that even the traditional view of f.o.b. pricing as a competitive process comes into question when invasion of another firm's submarket through price discrimination is considered (Greenhut, Norman, and Chao-shun Hung, 1987; Benson and Greenhut, 1989). Competitive spatial price discrimination enhances consumer welfare *relative* to a basing-point system, which welfare consequence is a rather clear theoretical prediction, not the unpredictable empirical issue suggested by Haddock.[6] Furthermore, BPP is a hybrid price system characterized by freight absorption and phantom freight, as all firms at sites other than the base point are obliged to price discriminate while those located at the base point price f.o.b. mill.

When a *true* basing-point pricing system arises, it is likely to have been imposed as a result of a cooperative process (Machlup, 1940; Stigler, 1949) under which the distant sellers feel obliged (coerced) to follow the established system (Greenhut, 1956; Wilcox, 1960). As such, it provides strong corroborative evidence of cooperative pricing, particularly when accompanied by organized enforcement efforts and punishment of those who cut price in violation of rate books, and so on. When Haddock's *quasi*-base-point price system arises, it is in the form of discounts offered to buyers located most proximate to a distant seller.

Two final issues warrant mention. (1) Haddock's statement (1990, p. 957) that "It seems peculiar to try to settle an essentially *empirical* issue through pure theory,..." (emphasis added) reveals a failure to appreciate the place of theory in understanding the world we live in. Specifically, what is the empirical issue? Surely not the issue of rivalrous versus collusive behavior. That issue is a purely theoretical one that can be resolved only through careful consideration of the theory of pricing behavior. A theoretical model explains the circumstances under which a firm will price f.o.b. or follow BPP. Then, and only then, does the empirical determination of the form of pricing, its circumstances *and* extent become relevant. (2) Based on theory, the reason for a distant firm's adherence to BPP requires just one restrictive condition: fear of the impacts that would follow from competitive pricing. On the other hand, the reasons why firms at the production center use BPP to protect rents rather than what otherwise would *appear* to be more profitable systems are the homogeneity/near homogeneity of their goods,

[5] For empirical evidence of this pricing *in* the United States, West Germany and Japan, see Greenhut (1981).

[6] Indeed, we find Haddock's discussion of welfare implications somewhat mystifying since the welfare benefits he discusses arise under NBPP (which is a form of spatial price discrimination) and BGN never suggested that such gains would be "modest." We are also surprised that Haddock turns to Austrian arguments to defend his position on competitive BPP that was originally based on a very non-Austrian static equilibrium model. Of course, producer surpluses are relevant, but the gains in consumer surplus from the breakdown of a basing-point system do not arise solely from a surplus transfer: the non-base point firm is also better off. Surely Haddock is not suggesting that BPP is acceptable because the base-point firms are better off even though consumers and distant firms lose out?

the simplicity of BPP, and the low costs in implementing, monitoring, and enforcing the system.

REFERENCES

Benson, Bruce L. and Greenhut, Melvin L., *American Antitrust Laws in Theory and in Practice*, Aldershot, England: Avebury, 1989.

______, ______, **and Norman, George,** "On the Basing-Point System," *American Economic Review*, September 1990, *80*, 584–88.

Burns, Arthur R., *The Decline of Competition*, New York: McGraw-Hill, 1936.

Greenhut, Melvin L., *Plant Location in Theory and Practice*, Chapel Hill: University of North Carolina Press, 1956; 4th printing, Westport, CN: Greenwood Press: 1982.

______, "Spatial Pricing in the U.S.A., West Germany and Japan," *Economica*, February 1981, *48*, 79–86.

______, **Norman, George and Hung, Chao-shun,** *The Economics of Imperfect Competition: A Spatial Approach*, Cambridge: Cambridge University Press, 1987.

Haddock, David D., "Basing-Point Pricing: Competitive vs. Collusive Theories," *American Economic Review*, June 1982, *72*, 289–306.

______, "On the Basing-Point System: Comment," *American Economic Review*, September 1990, *80*, 957–62.

Hoover, Edgar M., "Spatial Price Discrimination," *Review of Economic Studies*, 1936–37, *4*, 182–91.

Johnson, Ronald N. and Parkman, Allen M., "Spatial Monopoly, Non-zero Profits and Entry Deterrence: The Case of Cement," *Review of Economics and Statistics*, August 1983, *65*, 431–39.

Smithies, Arthur F., "Monopolistic Price Policy in a Spatial Market," *Econometrica*, January 1941, *9*, 63–73.

Stigler, George J., "A Theory of Delivered Price Systems," *American Economic Review*, December 1949, *39*, 1143–59.

______, "The Theory of Economic Regulation," *Bell Journal of Economics and Management Science*, Spring 1971, *2*, 3–21.

Thisse, Jacques F. and Vives, X., "On the Strategic Choice of Spatial Price Policy," *American Economic Review*, March 1988, *78*, 122–37.

Wilcox, Clair, *Public Policy Toward Business*, Rev. Ed., Homewood, IL: Richard D. Irwin, 1960.

Aetna Portland Cement Co.* v. *FTC, 157 F. 2nd, 1946, Respondents Brief, p. 127.

Annual Report of the Attorney General of the United States, Washington: U.S. Department of Justice, Office of the Attorney General, 1937.

Congressional Record: Proceedings and Debates of the Congress, Vol. 105, Washington: USGPO, 1959.

New York Times, February 20, 1939.

The Adjustment of Expectations to a Change in Regime: Comment

By RAYMOND P. H. FISHE AND MARK WOHAR*

In an article in this *Review*, Gregory Mankiw, Jeffrey Miron, and David Weil (1987, hereafter MMW) argue that "the founding of the Federal Reserve System in 1914 led to a substantial change in the behavior of nominal interest rates." They substantiate this claim by using data on three- and six-month time loan rates available at New York City banks between 1890 and 1933.

We take issue with their conclusions for two reasons. First, the data that they use are subject to error and do not always represent market transactions. There were many months in which no business was conducted in the time loan market, primarily because of financial panic or distress, and in these months only a "nominal" loan rate was reported. The nominal rate was arbitrarily set at the usury ceiling in New York, which was 6 percent over this period. Thus, for many months, the true market rate was unknown.

Second, their claim that World War I had little to do with the stochastic structure of interest rates is questionable. The outbreak of the war closed stock and bond exchanges in the United States and Europe for more than four months.[1] The closing of the markets because of the war precipitated a liquidity crisis. The birth of the Federal Reserve System helped mobilize resources to reduce this crisis, but the system's initial endowment of only $247 million was small in comparison to the $382 million in "emergency currency" that Secretary of the Treasury William McAdoo issued in response to the crisis.

The reopening of the bond market on November 28, 1914, and of the stock market on December 12, 1914, offered additional relief to the liquidity crisis by allowing investors and bankers to reorganize and value their portfolios in a formal market setting. Because of the many liquidity-creating actions instituted after the onset of War World I, it is not possible to claim that interest rates were more affected by the birth of the Federal Reserve System than, say, the issue of emergency currency or the reopening of the markets.

Our discussion of these points proceeds as follows. The data problems are documented in Section I. In Section II, we offer a simple solution to the problem of "missing" market loan rates. We find that the three-month series shows a break in late 1914 or early 1915, but that the six-month series shows a break in the middle of 1912. We argue that the actual break in the three-month series coincides more closely with the reopening of the major stock exchanges than the opening of the Federal Reserve System. A few brief closing remarks are offered in Section III.

I. Data Problems

The data examined by MMW were collected from two sources. One was the 1910 report of the National Monetary Commission to Congress, compiled by A. Piatt Andrew and entitled "Statistics for the United

*Department of Economics, University of Miami, Coral Gables, FL 33124; and Department of Economics, University of Nebraska, Omaha, NE 68182, respectively. We wish to thank Louis De Alessi, Al Holtmann, and two anonymous referees for helpful comments, as well as Cigdem Ural for providing research assistance. We assume responsibility for any remaining errors.

[1] On Monday, July 27, 1914, the Vienna and Brussels exchanges were closed. On Tuesday, the Paris Bourse and the Montreal and Toronto stock exchanges were closed. On Wednesday, stock and bond exchanges in Berlin, St. Petersburg, Amsterdam, Liverpool, Antwerp, Hamburg, and Frankfort suspended trading. The New York Stock Exchange and the London Stock Exchange closed on Friday, July 31, 1914.

States: 1867–1909." This volume contains weekly time loan rates from 1890 to 1909 taken from the *Commercial and Financial Chronicle* (*C&FC*), a popular business magazine of the day. The data from 1910 to 1933 were collected independently by MMW from the *C&FC*. For their empirical work, MMW created a monthly data series using time loan rates published in the first week of each month.

In the *C&FC*, most of the time loan rates are market rates, but there are periods where the data are not reported (nr), reported as nominal (nom), or reported with the provision that a commission (com) is paid to the lender. Andrew (p. 119) warned potential users of these data errors and omissions when he wrote: "It will be observed that at different times in the years prior to 1897 classified rates are lacking for time loans or for paper, or for both. These represent periods of disturbed conditions." A careful examination of the *C&FC* shows that this statement is also correct for a few years after 1897, particularly 1902, 1907, and 1918.[2] Table 1 provides a summary of these periods for the three- and six-month time loan rate series.[3]

It is not an accident that loan rates are quoted at 6 percent or 6 percent plus a commission during these problem periods. The state of New York set the usury rate at 6 percent between 1890 and 1933, and the National Banking Act required that all national banks abide by the usury laws of the state in which they operated.[4] National banks that violated usury ceilings could lose their charter, although none ever did. The comptroller of the currency was assigned the duty of monitoring national banks for usury ceiling violations. To avoid the usury limit, banks would charge a front-end commission to borrowers, but they would not quote an effective rate so as to avoid the unwanted attention of state authorities or the comptroller of the currency.

Because of these reporting inconsistencies, the data series collected by MMW does not reflect true market rates. In general, MMW underestimate the true market rate by failing to allow for commission payments. The important question, however, is whether these data errors will affect their claim that the beginning of the Federal Reserve marked the beginning of a new stochastic structure of interest rates. This issue is addressed in the next section.

II. New Estimates

We have chosen the simplest approach possible to correct for the errors inherent in the time loan data; that is, we have defined the largest time interval where the data are not censored to define our data series.[5] Be-

[2] Unfortunately, the *C&FC* is inconsistent in its treatment of commission payments. Sometimes it would note the amount of commission paid, so that an effective rate could be computed, and other times it would only report that commissions were involved. After 1905, however, it would generally, although not always, include commissions in its quoted rates. An examination of the *Wall Street Journal* during the period reveals a similar reporting pattern.

[3] There is a story behind each of the dates identified in Table 1. For example, time loan rates were quoted as 6 percent nominally for four months during 1918. During this period, the government undertook two bond issues to finance its participation in World War I. These issues, known as the third and fourth Liberty Loan issues, were subscribed to by banks that resold them to their customers. There was significant concern surrounding the success of these issues because the Germans had begun a major offensive against Allied positions in an attempt to win the war. To guarantee that the Liberty bonds sold, the Liberty Loan Committee, working with the Federal Reserve Board and the secretary of the treasury, pressured banks into limiting their issue of time loans. Accordingly, these time loan rates represent bids with almost no loans forthcoming at these rates.

[4] See John James (1978, p. 80) for a discussion of usury laws during this period. The actual statute is reprinted in James Cahill (1923), Article 25, Section 370. This statute exempts call loans over $5,000 from usury restrictions, but it does not exclude time loans.

[5] Other approaches to these data errors are possible, but they are not free of problems. For example, one may simply exclude the "bad" data from the series and reestimate MMW's model. This creates a problem because one does not know whether the excluded data define a break in the series, which is a possibility, particularly during the latter part of 1907 when call loan rates reached 125 percent while time loans were legally restricted to 6 percent. An alternative approach would be to develop a forecasting model to estimate

TABLE 1—DATA ERRORS FOR THREE- AND SIX-MONTH LOAN RATES

	Originally Reported (percent)		MMW Reported (percent)	
Date	Three-Month	Six-Month	Three-Month	Six-Month
Apr 12, 1890	nr	5.00	ex	ex
Apr 19, 1890	nr	5.00	ex	ex
Apr 26, 1890	nr	4.50	ex	ex
May 3, 1890	nr	5.00	4.50	5.00
May 10, 1890	nr	5.00	ex	ex
May 17, 1890	nr	5.00	ex	ex
May 24, 1890	nr	5.75	ex	ex
May 31, 1890	nr	5.00	ex	ex
Aug 16, 1890	6 nom	6 nom	ex	ex
Aug 23, 1890	6 nom	6 nom	ex	ex
Aug 30, 1890	6 nom	6 nom	ex	ex
Sep 6, 1890	6 nom	6 nom	6.00	6.00
Sep 13, 1890	6 nom	6 nom	ex	ex
Sep 20, 1890	6 nom	6 nom	ex	ex
Nov 15, 1890	6+com	6+com	ex	ex
Nov 22, 1890	6+com	6+com	ex	ex
Nov 29, 1890	6+com	6+com	ex	ex
Dec 6, 1890	6+com	6+com	6.00	6.00
Dec 13, 1890	6+com	6+com	ex	ex
Jan 3, 1891	nr	6.00	6.00	6.00
May 23, 1891	nr	6.00	ex	ex
May 30, 1891	5.75	nr	ex	ex
Jun 6, 1891	nr	6.00	5.75	6.00
Jun 13, 1891	nr	6.00	ex	ex
Nov 14, 1891	5.25	nr	ex	ex
Jan 2, 1892	4.00	nr	4.00	4.75
Sep 24, 1892	4.50	nr	ex	ex
Oct 15, 1892	6.00	nr	ex	ex
Dec 3, 1892	nr	6.00	5.00	6.00
Apr 29, 1893	6.00	nr	ex	ex
Jul 1, 1893	6+com	6+com	6.00	6.00
Jul 8, 1893	6+com	6+com	ex	ex
Jul 15, 1893	6+com	6+com	ex	ex
Jul 22, 1893	6+com	6+com	ex	ex
Jul 29, 1893	nr	6+com	ex	ex
Aug 5, 1893	6+com	6+com	6.00	6.00
Aug 12, 1893	6+com	6+com	ex	ex
Aug 19, 1893	6+com	6+com	ex	ex
Aug 26, 1893	6+com	6+com	ex	ex
Apr 20, 1895	nr	4.25	ex	ex
Jan 4, 1896	6 nom	6 nom	6.00	6.00
Jan 11, 1896	6 nom	6 nom	ex	ex
Jan 18, 1896	6 nom	6 nom	ex	ex
Jan 25, 1896	6 nom	6 nom	ex	ex
Feb 1, 1896	6 nom	6 nom	6.00	6.00
Feb 8, 1896	6 nom	6 nom	ex	ex
Aug 8, 1896	6 nom	6 nom	ex	ex
Aug 15, 1896	6.00	6+com	ex	ex
Aug 22, 1896	nr	6+2 com	ex	ex
Aug 29, 1896	6+2 com	6+2 com	ex	ex
Sep 5, 1896	6+1 com	6+2 com	6.00	6.00
Sep 12, 1896	6+1 com	6+2 com	ex	ex
Sep 19, 1896	6+com	6+com	ex	ex
Oct 10, 1896	6 nom	6 nom	ex	ex
May 26, 1899	nr	3.75	ex	ex
Oct 13, 1899	nr	6.00	ex	ex
Oct 27, 1899	nr	6.00	ex	ex
Dec 22, 1899	6 nom	6 nom	ex	ex

TABLE 1—(*Continued*)

	Originally Reported (percent)		MMW Reported (percent)	
Date	Three-Month	Six-Month	Three-Month	Six-Month
Jul 6, 1900	nr	4.25	3.25	4.25
Nov 9, 1900	nr	4.75	ex	ex
Feb 22, 1901	nr	4.00	ex	ex
Mar 21, 1902	nr	4.50	ex	ex
Mar 28, 1902	nr	4.50	ex	ex
Apr 4, 1902	nr	4.00	4.25	4.00
May 2, 1902	nr	4.25	4.50	4.25
May 9, 1902	nr	4.87	ex	ex
Sep 19, 1902	6+1 com	6+1 com	ex	ex
Sep 26, 1902	6+1 com	6+1 com	ex	ex
Oct 3, 1902	6+com	6.00	6.25	6.00
Oct 10, 1902	6+com	6+com	ex	ex
Oct 17, 1902	6+1 com	6.00	ex	ex
Nov 21, 1902	6+com	5.75	ex	ex
Dec 12, 1902	6+com	6.00	ex	ex
Jan 2, 1903	nr	6.00	5.25	5.25
Mar 6, 1903	nr	5.25	5.25	5.25
Apr 24, 1903	nr	4.67	ex	ex
May 22, 1903	nr	4.50	ex	ex
Oct 25, 1907	6.5 nom	6 nom	ex	ex
Nov 1, 1907	14.00	nr	14.00	6.00
Nov 8, 1907	13.50	nr	ex	ex
Nov 15, 1907	13.50	nr	ex	ex
Nov 22, 1907	13.50	nr	ex	ex
Nov 29, 1907	13.50	nr	ex	ex
Dec 6, 1907	nr	nr	10.00	7.00
Jul 9, 1909	2.37	nr	ex	ex
Aug 5, 1910	3.87	nr	3.62	5.00
Jul 31, 1914	6 nom	6 nom	ex	ex
Mar 1, 1918	6 nom	6 nom	6.00	6.00
Mar 8, 1918	6 nom	6 nom	ex	ex
Mar 15, 1918	6 nom	6 nom	ex	ex
Mar 22, 1918	6 nom	6 nom	ex	ex
Mar 29, 1918	6.00	6 nom	ex	ex
Apr 5, 1918	6.00	6 nom	6.00	6.00
Apr 12, 1918	6.00	6 nom	ex	ex
Apr 19, 1918	6.00	6 nom	ex	ex
Apr 26, 1918	6.00	6 nom	ex	ex
May 3, 1918	6.00	6 nom	6.00	6.00
May 10, 1918	6.00	6 nom	ex	ex
May 17, 1918	6.00	6 nom	ex	ex
May 24, 1918	6.00	6 nom	ex	ex
May 31, 1918	6.00	6 nom	ex	ex
Sep 6, 1918	6 nom	6 nom	6.00	6.00
Sep 13, 1918	6 nom	6 nom	ex	ex
Sep 20, 1918	6 nom	6 nom	ex	ex
Sep 27, 1918	6 nom	6 nom	ex	ex
Mar 11, 1933	nr	nr	ex	ex

Source: Andrew (pp. 119–38); *Commercial & Financial Chronicle*, 1909–33.
Notes: "ex" indicates that this week was excluded from the data collected by MMW; "nr" indicates that no rates were reported for this week; "nom" indicates that rates are quoted but no business is conducted; and "com" indicates that a commission is paid in addition to the legal rate of 6 percent.

TABLE 2—SWITCH DATES USING MONTHLY DATA FOR INTEREST RATE MODEL

$$r_{t+1} = \kappa + \rho r_t + \nu_{t+1}$$

	Three-Month Rate Jan 1908–Feb 1918		Six-Month Rate Sept 1910–Feb 1918	
Date	$-\text{Log } L$	Posterior Odds Ratio	$-\text{Log } L$	Posterior Odds Ratio
1912: 1	131.4	0.111	63.4	0.277
2	131.1	0.162	63.6	0.236
3	130.8	0.212	63.0	0.415
4	130.7	0.234	63.4	0.274
5	130.2	0.372	62.9	0.469
6	129.8	0.541	62.1	1.000
7	129.5	0.757	63.2	0.340
8	129.6	0.686	65.7	0.029
9	131.7	0.087	66.0	0.020
10	134.3	0.007	66.1	0.019
11	137.2	0.000	66.0	0.020
12	139.6	0.000	65.4	0.036
1914: 7	139.2	0.000	67.5	0.047
8	131.4	0.114	69.7	0.054
9	132.7	0.031	67.5	0.047
10	134.0	0.009	68.1	0.026
11	134.9	0.004	68.3	0.020
12	129.9	0.509	65.2	0.045
1915: 1	130.0	0.485	65.5	0.035
2	129.2	1.000	65.0	0.054
3	130.0	0.456	65.5	0.034
4	130.7	0.225	65.9	0.022
5	131.5	0.105	66.3	0.015
6	132.1	0.054	66.7	0.011
7	132.9	0.026	67.0	0.075

Notes: Log L is the log of the likelihood function. The posterior odds ratio is calculated assuming a diffuse prior.

cause MMW selected the first week in each month, their three-month loan rate data are error-free for months between December 6, 1907 and March 1, 1918. The six-month loan rate data are error-free for months between August 5, 1910 and March 1, 1918. Other than the week ending July 31, 1914, the weekly data that we have collected are also error-free over these intervals.

The monthly data may be analyzed between these dates to search for a structural break. Alternatively, the weekly data may be used if a proxy can be found for July 31, 1914. For this proxy, we use the first interest rate quoted during the following week. This proxy is quite reasonable because World War I effectively started on Friday, July 31, and the uncertainty concerning who would participate in the war continued into the following week.

Table 2 reports the results of our tests for a structural break using monthly data. We estimate the step-switching model described by MMW (p. 366), and our results are presented so that they are comparable to those

the "unobserved" rates. This would create a problem when testing for a structural break, however, because the forecasting model is assumed to be correct over all of the data. If a break is found, the forecasting model must then be reestimated and another search for a break undertaken. This procedure is continued until the forecasting model and the structural break results are consistent with each other. We did not select the latter procedure because it introduces forecasting error into our tests and is not necessary given that our sample covered the relevant period.

reported in table 6 of their paper.[6] The switch dates are found when the logarithm of the likelihood function is maximized or, because we report the negative of the log likelihood to be comparable to MMW, when it is minimized. The three-month interest rate series produces a switch date in February 1915, which is slightly later than that reported by MMW for the same regressions.[7] In contrast, the switchpoint for the six-month series is June 1912, which is more than two years *before* the Federal Reserve System began operations.[8]

Comparing the posterior odds ratios of the two periods shows that the June/July period during 1912 is a likely candidate for a break in the three-month series, but a similar claim may not be made for the period in late 1914 or early 1915 for the six-month series.[9] Although these results suggest that the structure of three-month interest rates changed near the founding of the Federal Reserve, it is clear that the entire term structure was not similarly affected. This is an important result because if the Federal Reserve System changed the stochastic behavior of interest rates permanently, as suggested by Gregory Mankiw and Jeffery Miron (1986), then one would expect that this change would be present in both the three-month *and* six-month rates.

Table 3 refines the estimation of these switch points by using weekly data on three- and six-month interest rates. The three-month series produces a switch point during the first week in December 1914, while the six-month series suggests that the switch point occurs during the second week of June 1912. Both of these results are in general agreement with the switch points found using the monthly data series, although the six-month switch point is precisely where it is expected to be whereas the three-month switch point is two months away from the date predicted using monthly data.

These findings offer only limited support for MMW's conclusion that there is a structural break in interest rates after the Federal Reserve began operations on November 16, 1914. The weekly results for the three-month series, which are the closest to supporting MMW's hypothesis, suggest that the break occurred sometime during the first week of December 1914, not in the week the Federal Reserve went into business. This presents a causal dilemma because there were other events affecting financial markets during the weeks before and after the opening of the Federal Reserve System, and these events may be responsible for the break in three-month rates.

The event that immediately preceded the structural break identified in Table 3 for the three-month series is the reopening of bond trading on the New York Stock Exchange. Both the bond and stock markets provided ready sources for liquidity during 1914. Before the onset of the war, the bond market traded nearly $61 million per month and the New York Stock Exchange averaged $574 million per month.[10] With the closing

[6] The models were also estimated with seasonal dummies included. The lack of degrees of freedom inhibited a complete search of the monthly and weekly samples, but when the regressions could be calculated, using all of the dummies, the switch point results were the same as those reported here.

[7] This is the second switch point found for the three-month data. The first switch point is August 1909. Because the first switch occurs very early in the data, we searched for a second switch point. Both switch points, however, are statistically significant, ($F_{4,121} = 13.98$).

[8] MMW do not report a switch date for the six-month series using the simple model with the rate lagged one period as the independent variable. Estimating this model for their data over the period 1890 to 1933 reveals that the switch occurred in June 1901.

[9] The switch point identified in 1912 may be due to the uncertain political climate at the time. President Taft was engaged in a tough battle for the Republican nomination with ex-President Roosevelt. The stock market appeared to favor the renomination of Taft, but the outcome of the fight was not known until the convention in June. After Governor Woodrow Wilson of New Jersey won the Democratic nomination on the 46th ballot, taken July 2, 1912, and adopted a "radical" platform, Roosevelt reentered the race as a third-party candidate. The markets were alarmed at the turn of events because Wilson's platform was viewed as anti-business.

[10] *CF&C*, Bank & Quotation Section, September 5, 1914, p. 19.

TABLE 3—SWITCH DATES USING WEEKLY DATA FOR INTEREST RATE MODEL

$$r_{t+1} = \kappa + \rho r_t + v_{t+1}$$

	Three-Month Rate Dec 13, 1907 to Feb 22, 1918			Six-Month Rate Aug 12, 1910 to Feb 22, 1918	
Date	$-\text{Log } L$	Posterior Odds Ratio	Date	$-\text{Log } L$	Posterior Odds Ratio
1914: Oct 2	217.9	0.013	1912: Apr 5	−16.4	0.046
Oct 9	216.5	0.054	Apr 12	−18.8	0.071
Oct 16	216.0	0.092	Apr 19	−17.4	0.123
Oct 23	216.4	0.059	Apr 26	−16.2	0.040
Oct 30	217.1	0.030	May 3	−16.6	0.055
Nov 6	217.8	0.015	May 10	−16.9	0.074
Nov 13	215.9	0.104	May 17	−17.4	0.127
Nov 20	214.7	0.330	May 24	−17.8	0.187
Nov 27	214.9	0.268	May 31	−18.3	0.325
Dec 4	213.6	1.000	Jun 7	−18.9	0.568
Dec 11	214.2	0.569	Jun 14	−19.5	1.000
Dec 18	214.7	0.325	Jun 21	−19.1	0.677
Dec 25	215.3	0.187	Jun 28	−19.3	0.859
Dec 31	215.8	0.107	Jul 5	−18.7	0.473
1915: Jan 8	216.5	0.053	Jul 12	−18.0	0.231
Jan 15	216.1	0.084	Jul 19	−17.2	0.105
Jan 22	216.2	0.072	Jul 26	−17.7	0.168
Jan 29	216.8	0.041	Aug 2	−16.9	0.077
Feb 5	217.5	0.020	Aug 9	−17.4	0.128
Feb 11	218.2	0.010	Aug 16	−17.9	0.221
Feb 19	219.9	0.005	Aug 23	−18.4	0.361
Feb 26	219.6	0.003	Aug 30	−17.2	0.102

Notes: Log L is the log of the likelihood function. The posterior odds ratio is calculated assuming a diffuse prior.

of these markets, investors had to find liquidity from other sources, primarily banks, thereby increasing interest rates. The bond market reopened immediately before the structural break identified for the three-month series in Table 3, on November 28, and immediately after this break, on December 12, the New York Stock Exchange reopened. Because of the volume of transactions involved, these events may be partly responsible for the rapid decrease in interest rates during December 1914 and January 1915 and thus may have precipitated the structural break in the three-month data series.[11]

Using the monthly figures above, the transaction volume that was loss due to the closing of the stock and bond markets totaled nearly $2.5 billion. This amount literally swamps the transactions that were facilitated by the Federal Reserve System in November and December of 1914. The Federal Reserve System began operations with $247 million in assets, of which about $228 million represented reserve deposits from member banks and $18 million represented paid-in capital. The Federal Reserve did not act to inject these resources into the economy during 1914. As Table 4 shows, the amount of bills discounted and loans

[11] If the reopening of the markets was important, then their closing is expected to be important too. We reestimated a switch point for the three-month data using the period June 21, 1912, to December 4, 1914. This period is just after the switch observed for the six-month data (the posterior odds ratio suggested the possibility of a switch in the three-month data around this date) and just before the switch observed for the three-month data. A significant break was found on the week ending July 31, 1914, which marks the beginning of World War I and the closing of the markets.

TABLE 4—ASSETS OF THE FEDERAL RESERVE SYSTEM, 1914
(1,000s)

Date	Total Cash	Bills Discounted and Loans	Total Assets	Ratio of Bills and Loans to Total Assets (percent)
Nov 20, 1914	$241,408	$5,626	$247,158	2.3
Nov 27, 1914	262,470	7,383	270,018	2.7
Dec 4, 1914	262,932	9,844	273,084	3.6
Dec 11, 1914	260,243	10,257	272,476	3.8
Dec 18, 1914	258,287	9,043	269,990	3.4
Dec 24, 1914	258,316	8,552	271,683	3.2
Dec 31, 1914	255,647	10,593	277,844	3.8

Source: Exhibit K, *Annual Report*, Federal Reserve Board, 1914.
Notes: Total cash includes gold coin and certificates, legal-tender notes, silver certificates, and subsidiary coin. The difference between total assets and cash plus bills and loans equals miscellaneous assets.

made represented less than 4 percent of the resources of the system. The Federal Reserve literally kept nearly 96 percent of its resources idle during crucial months at the beginning of World War I.[12]

Secretary of the Treasury William McAdoo did more to provide liquidity to the banking system and facilitate commerce than did the Federal Reserve System. With the stock exchanges closed, public panic led to hoarding of currency. By moving quickly in early August 1914, Secretary McAdoo, with the assistance of Congress, authorized the issue of $500 million in emergency currency as provided for by the Aldrich-Vreeland Act of 1908.[13] The cost of emergency currency was only 3 percent initially, whereas the New York clearinghouse charged 6 percent for clearinghouse loan certificates and the New York Federal Reserve bank charged 6 percent to discount loans with a maturity greater than 30 days.[14,15] It is no wonder then that the discount operations of the Federal Reserve were limited to about $10 million during 1914, while the U.S. Treasury had "sold" over $382 million in emergency currency, an amount that represented 10.2 percent of the existing money stock.[16]

Although one may debate whether or not the events discussed above resulted in a permanent change in the stochastic structure of interest rates, it is clear that they cast doubt on the claim that the founding of the Federal Reserve System was the primary factor.

III. Conclusions

The data on time loan rates used by MMW are subject to error, primarily because of usury laws and reporting inconsis-

[12] The performance of the Federal Reserve System was not much different during 1915. At the end of December 1915, only 12.1 percent of the system's assets were committed to bills and securities (Federal Reserve Board, *Banking and Monetary Statistics: 1914–1941*, p. 330).

[13] The emergency currency was identical to national bank notes issued after 1908, and thus was easily introduced into circulation. National banks found the emergency currency appealing because it met their demand for liquidity, which was hampered by public hoarding, and it was a cheap source of capital.

[14] The *Wall Street Journal*, August 12, 1914, commented on the issue of emergency currency at 3 percent: "'Emergency currency' under present conditions is a misnomer. What it really is, is a makeshift and temporary rediscount market."

[15] The discount rate at the New York Federal Reserve Bank was 6 percent from November 16, 1914, to December 18, 1914. On December 18, 1914, the discount rate dropped to 5.5 percent (Exhibit M, *Annual Report*, Federal Reserve Board, 1914).

[16] Comptroller of the Currency, *Annual Report*, 1915, Vol. I, p. 90.

tencies. Using a subset of these monthly data, which is error-free, and a weekly data set, we find that the six-month series does not support their claim that there was a regime change in 1914 and that the three-month series offers only superficial support. These findings suggest that the economic events during the period did not affect the term structure similarly.

A careful examination of the events surrounding the break point in the three-month series suggests that the reopening of the stock and bond markets and the issue of emergency currency may have affected the behavior of three-month rates. In addition, as an operating entity, the Federal Reserve was largely benign during 1914, using less than 4 percent of its resources to stimulate the economy. In total, these findings suggest that the birth of the Federal Reserve System was less important than MMW have claimed.

REFERENCES

Andrew, A. Piatt, *Statistics for the United States: 1867–1909*, National Monetary Commission, S.Doc 570, 61st Congress, 2d session, 1910.

Cahill, James C., *Cahill's Consolidated Laws of New York*, Chicago: Callaghan and Company, 1923.

Mankiw, N. Gregory, Miron, Jeffrey A. and David N. Weil, "The Adjustment of Expectations to a Change in Regime: A Study of the Founding of the Federal Reserve," *American Economic Review* June 1987, *77*, 358–74.

______ **and Miron, Jeffrey A.,** "The Changing Behavior of the Term Structure of Interest Rates," *Quarterly Journal of Economics*, May 1986, *101*, 221–28.

James, John A., *Money and Capital in Postbellum America*, Princeton, NJ: Princeton University Press, 1978.

Board of Governors of the Federal Reserve System, *Annual Report*, Washington: USGPO, 1914.

______, *Banking and Monetary Statistics: 1914–1941*, Washington: USGPO, 1976.

Commercial and Financial Chronicle, Bank and Quotation Section, 1909–1933.

Comptroller of the Currency, *Annual Report,* Washington: USGPO, 1914–15.

The Adjustment of Expectations to a Change in Regime: Reply

By N. Gregory Mankiw, Jeffrey A. Miron, and David N. Weil*

In their comment on our paper (1987), Raymond Fishe and Mark Wohar (1990) dispute our conclusion that the founding of the Federal Reserve caused a change in the monetary regime shortly after 1914. They make two unrelated points. First, they note that our interest rate data are contaminated by measurement error, suggesting that our results on the timing of the regime change may be an artifact of this measurement error. Second, assuming that we were correct that the change in regime occurred in late 1914 or early 1915, they argue that the Federal Reserve could not have been responsible for the change.

In this reply we first show that alternative ways of dealing with the measurement error problem, including the method proposed by Fishe and Wohar, do not substantially alter the results in our original paper. We then argue that their attribution of the change in regime to factors other than the founding of the Fed is implausible because the change in the behavior of interest rates was permanent, while these other factors were transitory. In the third section we examine an anomaly raised by the Fishe and Wohar results.

I. Measurement Error

Fishe and Wohar focus on the issue discussed by Jeffrey Miron (1989) that usury laws imposed a ceiling of 6 percent on the reported interest rate. We agree that this ceiling presents a problem with the data by making the reported interest rate differ from a true market interest rate for certain observations. In particular, for about 2 percent of the observations in our sample, our data include a reported interest rate of 6 percent when the true market interest rate was above 6 percent.[1]

When using the data to estimate the timing of the change in regime, there are alternative ways to treat the questionable observations. In our original paper, we simply included all the data. Because the data near to the estimated switch date do not suffer from this measurement error problem, there is no obvious reason to suspect the estimate of the switch date.

Fishe and Wohar suggest that one should restrict the sample to the longest stretch around the founding of the Fed in which there are no questionable observations. The problem with this method is that the remaining sample period is extremely short, implying that one is excluding a large number of perfectly good observations. Yet even using this procedure, Fishe and Wohar obtain essentially the same result that we originally reported for the switch date for the stochastic process followed by the three-month interest rate.

Another way to deal with measurement error is to exclude those observations for which the reported interest rate is suspect. This technique allows one to use most of the observations. When we implement this procedure for the three-month interest rate process, we find that the switch date is December 1914 when we exclude month dummies and February 1915 when we include month dummies. For the Modigliani-Sutch equation that relates the six-month rate and the three-month rate, the switch date is June 1915 when we exclude or in-

*Departments of Economics, Harvard University, Cambridge, MA, 02128; Boston University, Boston, MA, 02215; Brown University, Providence, RI 02912.

[1] When comparing the data we used with the data collected by Fishe and Wohar, we also discovered several coding errors in our data. These errors are usually very minor and do not influence any of our reported results. The corrected data are available upon request.

clude month dummies. These results are almost identical to those in our original paper.

None of these methods of dealing with the measurement error in the data is ideal. Yet the results we reported are robust to the alternative methods. There is no evidence that any of the results we reported are attributable to measurement error.

II. The Cause of the Change in Regime

Fishe and Wohar argue that the founding of the Fed was not the cause of the change in stochastic process driving interest rates. They point out that many other economic events, such as the outbreak of World War I, the closing and reopening of the bond and stock markets, and the issue of emergency currency, were occurring at about the same time as the opening of the Fed. They suggest that one of these other events may have been responsible for the change in the behavior of interest rates.

We recognized this possibility in our original paper, where we wrote:

> The year 1914 also saw the outbreak of World War I. Our estimates of the stochastic process of the short-term interest rate indicate that the short rate followed essentially the same process in the 1915–18 period as in the 1919–33 period. It appears, therefore, that the war was not itself the major factor of the regime change.

All of the events that Fishe and Wohar document were transitory in nature and therefore cannot explain a permanent change in regime.

An alternative explanation for the permanent change in regime is the abandonment of the Gold Standard, as suggested by Truman Clark (1986). We examined this possibility in Robert Barsky et al. (1988), where we discussed theoretical and empirical reasons to doubt the gold standard explanation. The resolution of this question, however, is not crucial to the central conclusion of our original paper. In that paper, we noted that "while our econometric results below point to the founding of the Fed rather than the abandonment of the Gold Standard as the likely cause of the regime change, our analysis of the adjustment of expectations does not rely on the Fed being the source of the change."

III. The Change in the Process of the Six-Month Rate

Fishe and Wohar point out that the switch date for the univariate process for the six-month interest rate is different from the switch dates for the three-month interest rate process and for the Modigliani-Sutch equation. This anomaly is present using our original data and using either their suggested correction for measurement error or the correction that we prefer.

All three techniques do produce local peaks in the likelihood function around the founding of the Fed. This finding leads us to believe that while one of the factors determining the stochastic process of the six-month rate (that is, the process for the three-month rate) did change when the Fed was founded, other determinants of the process for the six-month rate changed at other times. To make this point more concretely, consider the model for the six-month rate presented in our earlier paper,

$$(1) \qquad R_t = 1/2\,(1+\rho^3)r_t + \theta_t,$$

where R_t is the six-month rate, r_t is the three-month rate, ρ is the autoregressive parameter of the three-month rate, and θ_t is the term premium. Our analysis focused on whether there was a change in the coefficient on the short rate in this equation.

The univariate process for R_t changes when either the process for r_t or the process for θ_t changes. If these processes change at different times, and one estimates a switching regression for the univariate process for the long rate, either switch date may be found. The focus of our original paper was the behavior of the short rate and the relationship between the long and the short rate. We found that these two equations changed in late 1914 or early 1915. We interpreted the results as a change in the actual value and the public's perception of ρ. The finding that the univariate process for the six-month rate switched at a different point in time suggests that there also may have been changes in the process followed by θ_t. Yet this finding in no way vitiates our result that the behavior of the term structure changed to reflect the new process followed by the short rate around the time that the Fed was founded.

IV. Conclusion

We have learned two things by reflecting on Fishe and Wohar's comment on our 1987 paper. First, the results in our original paper on the timing of the change in the behavior of interest rates are not affected by the measurement error that arises because of state usury laws. Indeed, we are now even more confident in our results, because Fishe and Wohar have shown that using weekly rather than monthly data on three-month interest rates produces the same findings on the nature and timing of the regime change. Second, the univariate behavior of six-month interest rates appears anomalous when compared to the behavior of three-month interest rates and the behavior of the Modigliani-Sutch term structure equation relating six-month and three-month rates.

On the question of what caused the behavior of interest rates to change so dramatically in 1914, we remain convinced that the Fed is the most likely culprit. Since the elimination of transitory movements in interest rates in 1914, short-term interest rates have been close to a random walk (N. Gregory Mankiw and Miron, 1986), Because this behavior of interest rates has persisted throughout the Fed's 75-year history despite many other changes in economic policy, the Fed seems the most likely cause of the change in the behavior of interest rates.

REFERENCES

Barsky, Robert B., Mankiw, N. Gregory, Miron, Jeffrey A. and Weil, David N., "The Worldwide Change in the Behavior of Interest Rates and Prices in 1914," *European Economic Review*, June 1988, *32*, 1123–47.

Clark, Truman, "Interest Rate Seasonals and the Federal Reserve," *Journal of Political Economy*, February 1986, *94*, 76–125.

Fishe, Raymond P. H. and Wohar, Mark, "Did the Federal Reserve System Really Represent a Regime Change in 1914?" *American Economic Review*, September 1990, *80*, no. 4, 968–76.

Mankiw, N. Gregory and Miron, Jeffrey A., "The Changing Behavior of the Term Structure of Interest Rates," *Quarterly Journal of Economics*, May 1986, *101*, 211–28.

______, ______ and Weil, David N., "The Adjustment of Expectations to a Change in Regime: A Study of the Founding of the Federal Reserve," *American Economic Review*, June 1987, *77*, 358–74.

Miron, Jeffrey A., "The Founding of the Fed and the Destabilization of the Post-1914 U.S. Economy," in Marcello de Cecco and Alberto Giovannini, eds., *A European Central Bank? Perspectives on Monetary Unification after Ten Years of the EMS*, Cambridge: Cambridge University Press, 1989, 290–327.

Preliminary Announcement of the Program

ANNUAL MEETING
THE AMERICAN ECONOMIC ASSOCIATION

Washington, D.C., December 28–30, 1990

Thursday, December 27, 1990

10:00 A.M. Executive Committee Meeting

Friday, December 28, 1990

8:00 A.M. Savings and Investment Behavior in Developing Economies
Presiding: Terry Sicular, Harvard University
Papers: Catherine B. Hill, Williams College
A Precautionary Demand for Savings, Liquidity Constraints and Tests of the Permanent Income Hypothesis in Africa
Michael J. Athey and Prem S. Laumas, Northern Illinois University
Fragmented Capital Markets and Investment Spending: Evidence from a Developing Economy
Darryl McLeod and Parantap Basu, Fordham University
The Causes of Developing Country "Capital Flight": Some Cross-Section Evidence
Discussants: Christine Jones, World Bank

8:00 A.M. Policy Issues in Agricultural Economics
Presiding: Tim Phipps, West Virginia University
Papers: Lilyan Fulginiti, Iowa State University, and Richard Perrin, North Carolina State University
The Theory and Measurement of the Effects of Price Policies on Agricultural Productivity
Shangnan Shui, John C. Beghin, and Michael Wohlgenant, North Carolina State University
The Impact on the U.S. Cotton Industry of Removal of the Multiple Fiber Arrangement
Dean Lueck, Brigham Young University, and Terry Anderson, Montana State University
Property Rights in Indian Country: The Impact of Land Tenure on Agriculture
Discussants: Jeffrey Krautkraemer, Washington State University
Tim Phipps, West Virginia University
Katherine Reichelderfer, Resources for the Future

8:00 A.M. Immigration, Language and Ethnic Issues: Public Policy in Canada and the United States
Presiding: William T. Alpert, William H. Donner Foundation and University of Connecticut
Papers: Barry Chiswick, University of Illinois–Chicago, and Paul Miller, Queen's University
Language in the Labor Market: The Immigrant Experience in Canada and the United States
Gilles Grenier, University of Ottawa, and David Bloom, Columbia University
The Economic Status of Linguistic Ethnic Minorities: Hispanic-Americans and French-Canadians
Francois Vaillancourt, University of Montreal
Language and Public Policy in Canada and the United States
Discussants: William T. Alpert, William H. Donner Foundation and University of Connecticut
June O'Neill, Baruch College

8:00 A.M. Antitrust and Regulation
Presiding: Robert D. Willig, Princeton University and U.S. Department of Justice
Papers: Jonathan B. Baker, Dartmouth College
Econometric Analysis of Residual Demand
Robert D. Willig, Princeton University and U.S. Department of Justice
Merger Guidelines and Economic Theory
D. Mookherjee, Indian Statistical Institute, New Delhi, and I. P. L. Png, University of California–Los Angeles
Monitoring versus Investigation in Law Enforcement and Regulation

Discussants: JANUSZ ORDOVER, New York University
DANIEL SULLIVAN, Northwestern University
STEVEN SALOP, Georgetown University

8:00 A.M. POLITICAL-ECONOMIC DEVELOPMENTS IN THE BALKANS
Presiding: JOHN MICHAEL MONTIAS, Yale University
Papers: MARVIN R. JACKSON, JR., Arizona State University
The Bulgarian Economy After Zivkov
JOHN MICHAEL MONTIAS, Yale University
The Romanian Economy After the Genius of the Carpathians
KORI UDOVICKI, Yale University
Integration and Disintegration in Yugoslavia
Discussant: KEITH CRANE, Rand Corporation

8:00 A.M. SCOPE, SCALE, AND CAPITAL MEASUREMENT IN U.S. MANUFACTURING: RESULTS FROM THE LONGITUDINAL RESEARCH DATABASE
Presiding: ROBERT H. MCGUCKIN, U.S. Bureau of the Census
Papers: MARK E. DOMS, U.S. Bureau of the Census and University of Wisconsin
Measuring Vintage-Specific Depreciation Schedules Using Micro Longitudinal Data
MICHAEL GORT, U.S. Bureau of the Census and State University of New York-Buffalo
New Inputs, Old Inputs, and Productivity
MARY L. STREITWIESER, U.S. Bureau of the Census
Product Diversification, Economies of Scale and Scope in U.S. Manufacturing: Preliminary Findings of the Establishment and Firm Level
Discussants: CHARLES HULTEN, University of Maryland
ROBIN SICKLES, Rice University
JAMES MACDONALD, Rensselaer Polytechnic Institute

8:00 A.M. THE ECONOMICS OF SUSTAINABILITY
Presiding: JOEL DARMSTADTER, Resources for the Future
Papers: JOHN PEZZEY, University of Bristol
Sustainability and International Trade
MICHAEL TOMAN AND PIERRE CROSSON, Resources for the Future
Alternative Dimensions of Sustainability
Discussant: THOMAS TIETENBERG, Colby College

8:00 A.M. MARKET EQUILIBRIUM WITH MORAL HAZARD AND ADVERSE SELECTION
Presiding: PRAVEEN KUMAR, Carnegie Mellon University
Papers: THOMAS J. HOERGER, Vanderbilt University
Two-Part Pricing for Experience Goods and Services in the Presence of Moral Hazard and Adverse Selection
COLIN READ, University of Alaska
Landlords' Maintenance Strategies, Housing Quality, and Vacancies Under Imperfect Information
PRAVEEN KUMAR, Carnegie Mellon University
Optimal Product Innovation for a Durable-Goods Monopolist
Discussant: ANDREW S. JOSKOW, U.S. Department of Justice

8:00 A.M. POVERTY IMPACTS UNDER ADJUSTMENT LENDING (Joint Session with the American Committee on Asian Economic Studies)
Presiding: RICHARD HOOLEY, University of Pittsburgh
Papers: MICHAEL WALTON, World Bank
Analyzing Poverty Impacts under Adjustment
ERIK THORBECKE, Cornell University
Impact of Fiscal Retrenchment During Structural Adjustment on the Indonesian Socioeconomic System
PER PINSTRUP ANDERSEN, Cornell University
Effects of Changing Food Policies on the Poor in LDC's
Discussants: JERE BEHRMAN, University of Pennsylvania
MARK SUSSBERG, World Bank
ROMEO BAUTISTA, International Food Policy Research Institute

8:00 A.M. THE MARKET FOR CORPORATE CONTROL (Joint Session with the Association of Managerial Economists)
Presiding: MARK FEDENIA, University of Wisconsin-Madison
Papers: RONALD M. GIAMMARINO AND ROBERT L. HEINKEL, University of British Columbia
The Evolution of Firm Value and the Allocative Role of Greenmail
JAMES K. SEWARD AND JAMES P. WALSH, Dartmouth College
The Governance and Control of Voluntary Corporate Spinoffs: An Investigation of the Contracting Efficiency Hypothesis
JONATHAN M. KARPOFF AND PAUL H. MALATESTA, University of Washington
The Wealth Effects of Second Generation State Takeover Legislation
Discussants: BARTON L. LIPMAN, Carnegie Mellon University
KATHERINE SCHIPPER, University of Chicago
WAYNE MARR, Tulane University

8:00 A.M. LATIN IMMIGRATION AND THE U.S. ECONOMY IN THE 1990'S (Joint Session with Hispanic Professors of Economics and Business)
Presiding: JORGE SALAZAR-CARILLO, Florida International University
Papers: FRANCISCO RIVERA-BATIZ, Rutgers University-New Brunswick
Puerto Rican Migration and Its Impact
GILBERT CARDENAS, Pan-American University
Mexican Immigration and U.S. Labor Markets
ANTONIO JORGE AND RAUL MONCARZ, Florida International University
Cuban Immigration into the United States and Minority Business Development
Discussants: CORDELIA REIMERS, Hunter College
ROGER BETANCOURT, University of Maryland

10:15 A.M. WOMEN IN ECONOMICS
Presiding: NANCY M. GORDON, Congressional Budget Office
Papers: IVY E. BRODER, National Science Foundation and American University
New Evidence on Wage Differentials in Economics
REBECCA BLANK, Northwestern University
The Effect of Blind Refereeing in Economics
JUNE O'NEILL, Baruch College, AND NACHUM SICHERMAN, Rutgers University-New Brunswick
Is the Gender Gap in Economics Declining?
Discussants: CLAUDIA GOLDIN, Harvard University
ARLENE HOLEN, U.S. Office of Management and Budget
MARVIN KOSTERS, American Enterprise Institute

10:15 A.M. WHY ARE PRICES STICKY? EARLY RESULTS FROM AN INTERVIEW STUDY
Presiding: ROBERT M. SOLOW, Massachusetts Institute of Technology
Papers: ALAN S. BLINDER, Princeton University
Why Are Prices Sticky? Early Results from an Interview Study
Discussants: ROBERT SHILLER, Yale University
ROBERT J. GORDON, Northwestern University
ROBERT E. HALL, Stanford University
ROBERT M. SOLOW, Massachusetts Institute of Technology

10:15 A.M. THE MEASUREMENT OF INPUT QUALITY AND PRODUCTIVITY
Presiding: ERNST BERNDT, Massachusetts Institute of Technology and National Bureau of Economic Research
Papers: CHARLES R. HULTEN, University of Maryland and National Bureau of Economic Research
Embodied versus Disembodied Technical Change in Two-Digit U.S. Manufacturing Industries
LARRY ROSENBLUM, EDWIN DEAN, MARY JABLONSKI, AND KENT KUNZE, U.S. Bureau of Labor Statistics
Heterogeneous Labor Inputs and the Measurement of Productivity
ERNST R. BERNDT, Massachusetts Institute of Technology and National Bureau of Economic Research, CATHERINE J. MORRISON, Tufts University and National Bureau of Economic Research, AND DAVID O. WOOD, Massachusetts Institute of Technology
Assessing the Productivity of Information Technology Equipment in U.S. Manufacturing Industries
Discussants: MICHAEL INTRILIGATOR, University of California-Los Angeles
FINIS WELCH, University of California-Los Angeles
MARTIN N. BAILY, University of Maryland

10:15 A.M. Tort Law as a Regulatory System
Presiding: Christopher DeMuth, American Enterprise Institute
Papers: Susan Rose-Ackerman, Yale University
Tort Law in the Regulatory State
Robert Litan, Brookings Institution
The Impact of Tort Liability on Safety and Innovation: Lessons from the Scientific Community
Glenn Blackmon and Richard Zeckhauser, Harvard University
The Regulation of Automobile Liability Insurance: Lessons from Massachusetts
Discussants: Roger Noll, Stanford University
George Eads, General Motors Corporation
Robert Crandall, Brookings Institution

10:15 A.M. Diffusion of Development
Presiding: Edward Wolff, New York University
Papers: Richard R. Nelson, Columbia University
Diffusion of Development: Western Europe and North America
Marshall Goldman, Wellesley College and Harvard University
Diffusion of Development: Eastern Europe
Alice H. Amsden, New School for Social Research and Massachusetts Institute of Technology
Diffusion of Development: East Asia
Discussants: Paul David, Stanford University
Donald Harris, Stanford University
Henry Rosovsky, Harvard University

10:15 A.M. Social Values and the Transformation of Eastern Europe (Panel Discussion)
Presiding: Abram Bergson, Harvard University
Panel: Kenneth Arrow, Stanford University
Janos Kornai, Harvard University

10:15 A.M. Seminars on Recent Topics and Techniques*
Speaker: Jean Tirole, Massachusetts Institute of Technology
New Developments in Industrial Organization

10:15 A.M. Why is the Intellectual Debate on Macroeconomics Such a Shambles? (Roundtable)
Presiding: Francis M. Bator, Harvard University
Panel: Charles Schultze, Brookings Institution
John Taylor, Stanford University
John Berry, *Washington Post*
Alan Murray, *Wall Street Journal*

10:15 A.M. The Economics of Nuclear Power
Presiding: John L. Solow, University of Iowa
Papers: James Hewlett, U.S. Department of Energy
The Cost of Nuclear Power: Evidence from the United Kingdom
Mark McCabe and Richard Lester, Massachusetts Institute of Technology
Principals, Agents, and the Learning Curve: The Case of Nuclear Power Plant Construction
Geoffrey Rothwell and J. Bradford Jensen, Stanford University
Information Structures for Maintaining and Operating Nuclear Reactors
Discussants: Edward Kokkelenberg, State University of New York-Binghamton
Anthony Krautmann, DePaul University
John Marshall, University of California-Santa Barbara

10:15 A.M. The Role of Trade in North American Integration (Joint Session with the North American Economics and Finance Association)
Presiding: Mahmood A. Zaidi, University of Minnesota
Papers: Luis Ernest Derbez, World Bank
Trade Liberalization Policies in the North America Context
Harry P. Huizina, Stanford University, and K. C. Fong, University of California-Santa Cruz
The Free Trade Agreement and Union Wages in Canada
Colin I. Bradford, World Bank
A Global Trade and Growth Strategy for Correcting U.S. Imbalances
Discussants: Rudiger W. Dornbusch, Massachusetts Institute of Technology
George von Furstenberg, Indiana University
Richard B. Freeman, Harvard University

10:15 A.M. CONFLICT AND PEACE ECONOMICS II: DISARMAMENT EFFECTS AND NATIONAL PROTECTION (Joint Session with the Peace Science Society International)
Presiding: WALTER ISARD, Cornell University
Papers: GERALD F. ADAMS, University of Pennsylvania
Macroeconomic Effects of Disarmament in the Global Economy
JON HAVEMAN, ALAN V. DEARDORFF, AND ROBERT M. STERN, University of Michigan
The Economic Effects of Unilateral and Multilateral Reductions in Military Expenditures in the Major Industrialized and Developing Countries
WILLIAM G. SHEPHERD, University of Massachusetts
Efficient Levels of Post-Military Spending on National Prctection
Discussants: PETER H. PAULY, University of Toronto
SHERMAN ROBINSON, University of California-Berkeley

10:15 A.M. WHAT REMAINS OF COMPARATIVE ADVANTAGE? (Roundtable)
Presiding: DOMINICK SALVATORE, Fordham University
Papers: JAGDISH BHAGWATI, Columbia University
Level Playing Field and Comparative Advantage
PAUL KRUGMAN, Massachusetts Institute of Technology
Do New Trade Theories Require New Trade Policies?
WILLIAM BRANSON, Princeton University
Heckscher-Ohlin is Alive and Well and Living in the Pacific!
ROBERT BALDWIN, Univeristy of Wisconsin
Where Do We Stand on Comparative Advantage?

12:30 P.M. AEA/AFA JOINT LUNCHEON
Presiding: THOMAS C. SCHELLING, University of Maryland
Speaker: (TO BE ANNOUNCED)

2:30 P.M. PATENT RACES AND TECHNOLOGY TRANSFER
Presiding: DEBRA ARON, Northwestern University
Papers: REIKO AOKI, Ohio State University
R&D Competition for Product Innovation: A Race Without End
RICHARD JENSEN, University of Kentucky, AND MARIE THURSBY, Purdue University
Patent Races, Product Standards, and International Competition
BETH ANNE TERCEK, Boston College
North–South Technology Transfer in the Context of International Returns to Scale
Discussants: JOSEPH E. STIGLITZ, Stanford University
JENNIFER F. REINGANUM, University of Iowa
PAUL ROMER, Center for Advanced Studies in Behavioral Sciences

2:30 P.M. STABILIZATION AND REFORM IN EASTERN EUROPE AND LATIN AMERICA
Presiding: KENNETH FROOT, Massachusetts Institute of Technology
Papers: RONALD MCKINNON, Stanford University
Financial Control During the Transition from a Centrally Planned to a Market Economy
JOHN COCHRANE, University of Chicago, AND BARRY W. ICKES, Pennsylvania State University
Stopping Inflation in Reforming Socialist Economies
SEBASTIAN EDWARDS, University of California-Los Angeles
Stabilization Experience in Latin America: Lessons for Eastern Europe
THOMAS WOLF, International Monetary Fund
Stabilization and Adjustment in Eastern Europe
Discussants: RUDIGER DORNBUSCH, Massachusetts Institute of Technology
ELIANA CARDOSO, Tufts University
JAN SVEJNAR, University of Pittsburgh
LAWRENCE SUMMERS, Harvard University

2:30 P.M. CAN SANCTIONS SAVE FREE TRADE?
Presiding: C. MICHAEL AHO, Council on Foreign Relations
Papers: THOMAS O. BAYARD, Institute for International Economics
Reciprocity and Retaliation: Should Might Make Right?
MARK LEVINSON, *Journal of Commerce*
Can Sanctions Save Free Trade?
KENNETH OYE, Swarthmore College and Massachusetts Institute of Technology
Economic Discrimination and Political Effectiveness: In Defense of Bilateral Sanctions

Discussants: AVINASH DIXIT, Princeton University
C. MICHAEL AHO, Council on Foreign Relations

2:30 P.M. THE DEMAND FOR MONEY REVISITED
Presiding: JAMES BOUGHTON, International Monetary Fund
Papers: DAVID LAIDLER, University of Western Ontario
The Demand for Money: An Empirical Overview
BENNETT MCCALLUM, Carnegie Mellon University
The Demand for Money: A Theoretical Overview
JAMES BOUGHTON, International Monetary Fund
The Demand for Money: International Comparisons
Discussants: STEPHEN GOLDFIELD, Princeton University
DALE HENDERSON, Federal Reserve Board
RAY FAIR, Yale University

2:30 P.M. CHINESE ECONOMIC REFORMS, 1979–1989: LESSONS FOR THE FUTURE
Presiding: GREGORY C. CHOW, Princeton University
Papers: ROGER H. GORDON, University of Michigan
Chinese Enterprise Behavior under the Reforms
BARRY NAUGHTON, University of California-San Diego
Why Has Economic Reform Led to Inflation?
TERRY SICULAR, Harvard University
Moving Toward Market Allocation in Agriculture: Potential Problems
Discussants: LAWRENCE J. LAU, Stanford University
RICHARD D. PORTES, Birkbeck College
GERSHON FEDER, World Bank

2:30 P.M. TEACHING COLLEGE ECONOMICS
Presiding: STEPHEN BUCKLES, Joint Council on Economic Education
Papers: JOHN SIEGFRIED, Vanderbilt University, ROBIN BARTLETT, Denison University, W. LEE HANSEN, University of Wisconsin-Madison, ALLEN C. KELLEY, Duke University, DONALD MCCLOSKEY, University of Iowa, AND THOMAS TIETENBERG, Colby College
The B^- Economics Major: Can and Should We Do Better?
WILLIAM BECKER, Indiana University, ROBERT HIGHSMITH, Joint Council on Economic Education, PETER KENNEDY, Simon Fraser University, AND WILLIAM WALSTAD, University of Nebraska
A New Agenda for Research on Teaching College Economics
PHILLIP SAUNDERS, Indiana University
The Newly Revised Test of Understanding College Economics
Discussants: DAVID COLANDER, Middlebury College
CLAUDIA GOLDIN, Harvard University
ALAN BLINDER, Princeton University
ERIC HANUSHEK, University of Rochester

2:30 P.M. LABOR SUPPLY
Presiding: DAVID BLOOM, Columbia University
Papers: SHELLY LUNDBERG, Princeton University, AND ROBERT PLOTNICK, University of Washington
Teenage Childbearing Decisions: Do "Opportunity Costs" Matter?
LORI BOLLINGER, University of Pennsylvania
Diffusion and Adoption of Contraceptive Technology in Developing Countries
MANOUCHEHR MOKHTARI, University of Maryland-College Park, AND PAUL R. GREGORY, University of Houston
Backward Bends, Quantitative Constraints, and Labor Supply: Evidence from the Soviet Interview Project
FINIS WELCH, Unicon Research Corporation, AND KEVIN MURPHY, University of Chicago
Wages and Participation in the 1980's
Discussants: SANDERS KORENMAN, Princeton University
DAVID LOAM, University of Michigan
GARY BURTLESS, Brookings Institution
DAVID BLOOM, Columbia University

2:30 P.M. SMOKING, NUTRITION, AND HEALTH
Presiding: NANCY GORDON, Congressional Budget Office
Papers: GERARD RUSSO, University of Hawaii
The Demand for Physicians' Services and the Price of Cigarettes

W. Kip Viscusi, Duke University
Perception of Smoking Risks and Cigarette Smoking Behavior
Pauline M. Ippolito and Alan D. Mathios, Federal Trade Commission
Information, Advertising, and Health Choices: A Study of the Cereal Market
Discussants: Janet B. Mitchell, Center for Health Economics
Willard G. Manning, University of Michigan
Shelley I. White-Means, Memphis State University

2:30 P.M. Options and Bubbles: Employing Emerging Methodologies to Examine Questions in Financial Economics (Joint Session with the Economic Science Association)
Presiding: John Dickhaut, University of Minnesota
Papers: John O'Brien and Sanjay Srivastava, Carnegie Mellon University
Writing Options in the Laboratory
Hersh M. Shefrin, University of Santa Clara
Prospect Theory as a Basis for Examining Behavior with Call Options
Vernon Smith, University of Arizona
Bubble Crashes in the Laboratory: A Summary of What We Know Plus Some New Evidence
Discussants: Chester Spatt, Carnegie Mellon University
Daniel Friedman, University of California-Santa Cruz

2:30 P.M. Maintaining Infrastructure and Fiscal Credibility in Urban America (Joint Session with the National Tax Association)
Presiding: Roy Bahl, Georgia State University
Papers: Michael E. Bell, Johns Hopkins University
Infrastructure and Environmental Quality Requirements for Sustainable Growth
Helen F. Ladd, Duke University and Lincoln Institute of Land Policy
Fiscal Implications of Growth and Decline in U.S. Cities
Robert A. Bohm and Deborah Vaughn Nestor, University of Tennessee
Characteristics of the Solid Waste Crisis and Incentives to Create Capacity
Discussants: George Plesko, Northeastern University
Sally Wallace, U.S. Department of the Treasury
John M. Quigley, University of California-Berkeley

4:45 P.M. Richard T. Ely Lecture
Presiding: Thomas C. Schelling, University of Maryland
Speaker: George A. Akerlof, University of California-Berkeley
Procrastination, Indoctrination, and Obedience

Saturday, December 29, 1990

8:00 A.M. Gender and Productivity
Presiding: Francine Blau, University of Illinois-Urbana-Champaign
Papers: Solomon Polachek, State University of New York-Binghamton
An Analysis of Recent Trends in the Male–Female Wage Gap
Joni Hersch, University of Wyoming
Gender Differences in Wages: The Role of Human Capital, Working Conditions, and Housework
John Mullahy, Trinity College, and Jody L. Sindelar, Yale University
Substance Abuse and Gender Differences in Productivity
Discussants: Francine Blau, University of Illinois-Urbana-Champaign
Laurie Bassi, Georgetown University
Jean Mitchell, Florida State University

8:00 A.M. Effects of Taxation
Presiding: Jane G. Gravelle, Congressional Research Service
Papers: William M. Shobe, University of North Carolina
The Effects of Taxation on Capital Gains Realizations: Theory and Evidence
William Gale, University of California-Los Angeles, and John Karl Scholz, University of Wisconsin-Madison
The Effects of IRA's on Household Savings
James Andreoni and John Karl Scholz, University of Wisconsin–Madison
An Econometric Analysis of Charitable Giving and Interdependent Preferences

Discussants: LEONARD BURMAN, Congressional Budget Office
BARBARA MANN, U.S. Department of the Treasury
DANIEL FEENBERG, National Bureau of Economic Research

8:00 A.M. ENERGY INVESTMENTS IN DEVELOPING COUNTRIES: ENVIRONMENTAL ISSUES AND POLICY APPROACHES
Presiding: CORAZON M. SIDDAYAO, World Bank
Papers: MARCIA GOWEN, Winrock
Investments in Biomass Fuels: Environment and Policy Issues
CAROL DAHL, Louisiana State University
Energy Demand Elasticities in the Developing World and Their Implications for Global Environmental Issues
RALPH W. HUENEMANN, University of Victoria
Cost/Benefit Analysis of a Power Project in the Three Gorges Dam in China: Environmental Implications of a Persistent Error
GUNTER SCHRAMM, World Bank
Incorporating Uncertainty in Addressing Environmental Issues in Power Section Investments
JOHN SHEERIN, U.S. Department of State, KIRK R. SMITH, East–West Center, AND CORAZON M. SIDDAYAO, World Bank
The Global Warming Issue: Policy Approaches to Developing Countries' Problems
Discussants: MOHAN MUNASINGHE, World Bank
UZIEL NOGUEIRA, Inter-American Development Bank

8:00 A.M. ALTRUISM: EMPIRICAL STUDIES
Presiding: ROBERT LINDSAY, New York University
Papers: ROBERT H. FRANK, Cornell University
Are Economists Good Citizens?
JOHN CULLIS AND ALAN LEWIS, University of Bath, United Kingdom
United Kingdom Ethical Investments: A Case Study
CHARLOTTE D. PHELPS, Temple University
Money, Love, and Happiness
Discussants: MICHAEL WALDMAN, University of California–Los Angeles
RANDOLPH WESTERFIELD, University of Southern California
SHOSHONA A. GROSSBARD-SHECHTMAN, San Diego State University
PAUL L. WACHTEL, City University of New York

8:00 A.M. ECONOMICS IN SPACE
Presiding: LINDA COHEN, University of California-Irvine
Papers: MOLLIE MACAULEY AND MICHAEL TOMAN, Resources for the Future
Eye in the Sky: The Economics of Remote Sensing
HARVEY LEVIN, Hofstra University
Joint Ventures, Acquisitions, and Market-Type Transactions in Trading Orbit Spectrum Assignments in the Space Satellite Industry
NEIL DOHERTY, University of Pennsylvania
Risk-Bearing and Moral Hazard in Space
LINDA COHEN, University of California-Irvine, SUSAN EDELMAN, Columbia University, AND ROGER NOLL, Stanford University
American Commercial Aerospace Transport: A Political Economic Perspective
Discussants: ARTHUR DEVANY, University of California-Irvine
NANCY ROSE, Massachusetts Institute of Technology
JAMES DEARDEN, Lehigh University

8:00 A.M. ALCOHOL AND PUBLIC POLICY
Presiding: JOHN MULLAHY, Trinity College and Resources for the Future
Papers: FRANK J. CHALOUPKA, University of Illinois-Chicago and National Bureau of Economic Research, MICHAEL GROSSMAN, City University of New York Graduate Center and National Bureau of Economic Research, AND HENRY SAFFER, Kean College of New Jersey and National Bureau of Economic Research
Alcohol, Regulation, and Motor Vehicle Mortality
PHILIP J. COOK AND MICHAEL J. MOORE, Duke University
Drinking and Earnings
JODY SINDELAR, Yale University, AND JOHN MULLAHY, Trinity College and Resources for the Future
Alcohol Abuse, Alcohol Dependence, and Productivity

Discussants: Charles Phelps, University of Rochester
Willard G. Manning, University of Michigan
Paul Gertler, Rand Corporation

8:00 A.M. Local Public Finance
Presiding: John Yinger, Syracuse University
Papers: Robert M. Schwab and Wallace E. Oates, University of Maryland
Community Composition and the Provision of Local Public Goods: A Normative Analysis
Dilip Bhattacharyya and Robert W. Wassmer, Wayne State University
Taxation and Expenditure Decisions of Elected Local Officials
Charles A. M. de Bartholome, New York University
The Fiscal Effect of Community Composition on Public Expenditure and Welfare
Discussants: Gary J. Reid, University of Southern California
David J. Sjoquist, Georgia State University

8:00 A.M. U.S.–Korea Trade Relations (Joint Session with the Korea-America Economic Association)
Presiding: E. Kwan Choi, Iowa State University
Papers: Joosung Jun, Yale University
Tax Policy and Korean Direct Investment in the U.S.
Danny M. Leipziger, World Bank
New Issues in Korean Trade Policy
Won W. Koo, North Dakota State University
Trade Liberalization for Agricultural Products in Korea and Its Implications on U.S. Agricultural Exports
Marcus Noland, Institute for International Economics
The Impact of Prospective Changes in the Korean Trade Pattern on the U.S. Economy
Chong Liew, University of Oklahoma
The Effects of Labor Disputes on Korea–U.S. Trade Structure
Discussants: Marcus Noland, Institute for International Economics
Joosung Jun, Yale University
E. Kwan Choi, Iowa State University
Kei Mu Lee, Rice University
Won W. Koo, North Dakota State University

8:00 A.M. Current Development of Chinese Financial Management and Planning (Joint Session with the Society on Economics and Management in China)
Presiding: Cheng-few Lee, Rutgers University-New Brunswick
Papers: Janson Chang and Bob C. Li, World Bank
Enterprise Investment and Financing Policy Review for Guangdong Agriculture Reclamation
Myron J. Gordon, University of Toronto
The Agency Problem and Its Solution in China's State Enterprises
Cheng-few Lee, Rutgers University-New Brunswick, and Gili Yen, National Central University
Financial and Economic Relationship Between China and Taiwan
Discussants: D. Gale Johnson, University of Chicago
Frank C. Jen, State University of New York-Buffalo
Charles Ou, U.S. Small Business Administration

8:00 A.M. Reassessing the Role of Finance in Development (Joint Session with the Union for Radical Political Economists)
Presiding: Carmen Diana Deere, University of Massachusetts-Amherst
Papers: Ben Crow, Open University
Credit Conditions and the Context of Credit: The Case of the Bangladesh Grain Trade
Keith Griffin, University of California-Riverside
Foreign Finance and Domestic Savings: What Have Twenty Years Taught Us?
Laurence Harris, Open University
Public Finance and Financing Development
Discussants: Stanley Fischer, World Bank

10:15 A.M. Theoretical IO with Applications to R&D
Presiding: Elizabeth Hoffman, University of Arizona
Papers: Beth Allen, University of Pennsylvania
Choosing R&D Projects: An Informational Approach
Karen Palmer, Resources for the Future
Diversification by Regulated Monopolies and Incentives for R&D

Suzanne Scotchmer, University of California-Berkeley
Cooperation in R&D and the Breadth of Patent Protection
Discussants: Tara Vishwanath, Northwestern University
Paula-Ann Cech, Northwestern University
Esther Gal-Or, University of Pittsburgh

10:15 A.M. The Political Economy of Capital Mobility
Presiding: Eric Nilsson, Tufts University
Papers: Timothy Koechlin, Skidmore College
The Responsiveness of Domestic Investment to Foreign Economic Conditions: An Analysis of Seven OECD Countries
Gerald Epstein and Herbert Gintis, University of Massachusetts-Amherst
An Asset Balance Model of International Capital Market Equilibrium
Gary Dymski, University of Southern California, and Manuel Pastor, Occidental College
Misleading Signals: Bank Lending and the Latin American Debt Crisis
Discussants: Richard MacIntyre, University of Rhode Island
Jeffrey A. Frankel, University of California-Berkeley
John Williamson, Institute for International Economics

10:15 A.M. Economics of Growth and Stagnation
Presiding: Badiul A. Majumdar, Washington State University
Papers: Deepak K. Lal, University College, London
The Political Economy of Poverty, Equity, and Growth in 21 Developing Countries—A Summary of Findings
Badiul A. Majumdar, Washington State University
Economics of Growth and Stagnation: Korea Versus Bangladesh
Gustav F. Papanek, Boston University
Growth, Equity, and Politics: Contrasting Development Strategies in India and Pakistan
Discussants: Anne O. Krueger, Duke University
E. Wayne Nafziger, Kansas State University
Douglas O. Walker, United Nations

10:15 A.M. Transfers and Exchanges Outside the Marketplace
Presiding: Oded Stark, Harvard University and Bar-Ilan University
Papers: James Andreoni and John Karl Scholz, University of Wisconsin-Madison
An Econometric Analysis of Charitable Giving with Interdependent Preferences
Donald Cox, Boston College, and Emmanuel Jimenez, World Bank
Motives for Private Transfers over the Life Cycle
Theodore Bergstrom, University of Michigan
Systems of Benevolent Utility Interdependence
Oded Stark, Harvard University and Bar-Ilan University
Nonmarket Transfers: The Role of Altruism
Discussants: Kenneth J. Arrow, Stanford University
Gary S. Becker, University of Chicago

10:15 A.M. Path-Dependence in Economics: The Invisible Hand in the Grip of the Past
Presiding: Paul A. David, Stanford University
Papers: Steven Durlauf, Stanford University
Persistence and Multiple Equilibria in Aggregate Fluctuations
James Heckman, Yale University
Identifying the Hand of the Past: Distinguishing State Dependence from Heterogeneity
Paul R. Krugman, Massachusetts Institute of Technology
History as a Determinant of Location and Trade
Paul Milgrom, Stanford University
Complementarities and the Evolution of Manufacturing Organizations
Discussants: W. Brian Arthur, Stanford University
Paul A. David, Stanford University
Paul Romer, University of California-Berkeley
Robert M. Solow, Massachusetts Institute of Technology

10:15 A.M. Reform, Regulation, and the Economic Behavior of Socialist Firms
Presiding: Josef Brada, Arizona State University
Papers: Istvan Abel, Karl Marx University, Budapest, and John P. Bonin, Wesleyan University

JANEZ PRASNIKAR, University of Ljubljana, Yugoslavia, AND JAN SVEJNAR, University of Pittsburgh
The Objectives and Employment Behavior of Yugoslav Firms
NADEZHDA MIKHEEVA, Academy of Sciences USSR, AND JUDITH THORNTON, University of Washington
Structural Change and Integration of the Soviet Far East into the Pacific Rim Economy
Discussants: ISVAN DOBOZI, Institute for World Economy, Budapest, and Colorado School of Mines
PETER MURRELL, University of Maryland

10:15 A.M. PATTERNS OF MINORITY PROGRESS
Presiding: MARGARET C. SIMMS, Joint Center for Political Studies
Papers: HARRIET ORCUTT DULEEP, U.S. Commission on Civil Rights, AND SETH SANDERS, University of Chicago
Accounting for the Narrowing Wage Gap Between American-Born Asians and Whites
NADJA ZALOKAR, U.S. Commission on Civil Rights
Sources of Black Women's Economic Progress Since 1940
HARRIET ORCUTT DULEEP AND MARK C. REGETS, U.S. Commission on Civil Rights
Selectivity and Changes in the Relative Earnings of Hispanics: Evidence from Diverse Groups
JUNE O'NEILL, Baruch College
Educational Achievement and Racial Differences in Economic Status
Discussants: WALTER Y. OI, University of Rochester
WAYNE VROMAN, Urban Institute

10:15 A.M. SEMINARS ON RECENT TOPICS AND TECHNIQUES*
Speaker: JAMES POTERBA, Massachusetts Institute of Technology
New Developments in Public Finance

10:15 A.M. THE TAX REFORM ACT OF 1986: APPRAISAL AND PROSPECTS
Presiding: HENRY J. AARON, Brookings Institution
Papers: JANE G. GRAVELLE, Congressional Research Service
The Tax Reform Act of 1986 and Tax Equity
JOEL SLEMROD, University of Michigan
Did the Tax Reform Act of 1986 Simplify Matters?
GARY BURTLESS AND BARRY BOSWORTH, Brookings Institution
The Supply Side to Tax Reform: Saving, Investment and Labor Supply
Discussants: J. GREGORY BALLENTINE, KPMB Peat Warwick

10:15 A.M. THE CONTRIBUTIONS OF FRIEDRICH A. HAYEK (Joint Session with the History of Economics Society)
Presiding: BRUCE J. CALDWELL, University of North Carolina-Greensboro
Papers: BRUCE J. CALDWELL, University of North Carolina-Greensboro
Hayek's Methodology
MEGHNAD DESAI, London School of Economics
Hayek's Early Contributions to Economic Theory
WILLIAM BUTOS, Trinity College
Hayek's Theory of Psychology
JEREMY SHEARMUR, Institute for Human Studies
Hayek's Political Economy
KAREN VAUGHN, George Mason University
Hayek's Theory of Culture

10:15 A.M. ECONOMIC CONSEQUENCES OF RAISING THE MINIMUM WAGE (Joint Session with the Industrial Relations Research Association)
Presiding: FRANCINE D. BLAU, University of Illinois-Urbana-Champaign
Papers: DAVID CARD, Princeton University
Using Regional Variation in Wages to Measure the Effect of Minimum Wages
ALIDA CASTILLO-FREEMAN, National Bureau of Economic Research, AND RICHARD B. FREEMAN, Harvard University
The Effect of the U.S. Minimum Wage: The United States and Puerto Rico
LAWRENCE F. KATZ, Harvard University, AND ALAN B. KRUEGER, Princeton University
Does the Minimum Wage Improve Minimum Wage Jobs?
BRADLEY R. SCHILLER, American University
Training and Advancement of Minimum Wage Workers
Discussants: CHARLES C. BROWN, University of Michigan
LISA M. LYNCH, Massachusetts Institute of Technology

2:30 P.M. EMPIRICAL ANALYSES OF R&D AND PRODUCTIVITY GROWTH
Presiding: BRONWYN HALL, University of California-Berkeley
Papers: LINDA M. EDWARDS AND BETSY FIELD-HENDREY, Quens College, City University of New York
Unions and Productivity in the Public Sector: The Case of Solid Waste Collection
JANET W. TILLINGER, Texas A&M University
Dividend Taxation and Dividend Signalling: An Analysis of the Effects on Investment in Research and Development
SARAH J. LANE, Boston University
New Diggins in Bituminous Coal—The Structure of the Output Market as a Determinant of Investment in New Technology
CATHERINE MORRISON, National Bureau of Economic Research
The Impacts of Markups, Capacity Utilization, and Scale Economies on Productivity Growth: A Reevaluation Using Productive Theory
Discussants: LISA M. LYNCH, Massachusetts Institute of Technology
EDWIN MANSFIELD, University of Pennsylvania

2:30 P.M. DIRECT INVESTMENT IN THE UNITED STATES: WHAT ARE THE ISSUES?
Presiding: MARTIN KOHLI, New York State ICC
Papers: NORMAN J. GLICKMAN, Rutgers University-New Brunswick, AND DOUGLAS P. WOODWARD, University of South Carolina
Strategy and Structure of Japanese Manufacturing Investment
PAUL R. KRUGMAN, Massachusetts Institute of Technology, AND EDWARD M. GRAHAM, Institute for International Economics
Discussants: ROBERT KUTTNER, *The New Republic*
ALICE H. AMSDEN, New School for Social Research
LEE SMITH, New York State ICC

2:30 P.M. CAN FEMINISM FIND A HOME IN ECONOMICS?
Presiding: ROBERT POLLAK, Washington University and University of Pennsylvania
Papers: PAULA ENGLAND, University of Arizona
How Should Feminism Change Economic Theory?
CLAUDIA GOLDIN, Harvard University
The Pollution Theory of Discrimination
DIANA STRASSMAN, Rice University
Feminism and Economic Knowledge
NANCY FOLBRE, University of Massachusetts
The Male Domain of Reason: Androcentrism in Classical Political Economy
Discussants: REBECCA BLANK, Northwestern University
PAULETTE OLSEN, Wright State University

2:30 P.M. NEW RESEARCH ON THE UNDERCLASS
Presiding: ISABEL V. SAWHILL, Urban Institute
Papers: RONALD B. MINCY, Urban Institute
Where the Underclass Lives: Evidence from Cross-Sectional Data
EDWARD M. GRAMLICH, University of Michigan
Geographic Mobility and the Underclass
JONATHAN CRANE, Harvard University
Economic Models of Noneconomic Motivations for Underclass Behavior
Discussants: GARY BURTLESS, Brookings Institution
ROBERT H. HAVEMAN, University of Wisconsin-Madison
ROBERT D. REISCHAUER, Congressional Budget Office

2:30 P.M. THE ECONOMIC IMPACT OF RELIGIOUS FUNDAMENTALISM
Presiding: FREDERIC L. PRYOR, Swarthmore College
Papers: DEEPAK LAL, University College, London
The Economic Impact of Hindu Fundamentalism
TIMUR KURAN, University of Southern California
Has Islamic Economics Made a Difference?
LAURENCE R. IANNACCONE, University of Santa Clara
Heirs to the Protestant Ethic? The Economics of American Fundamentalists
Discussants: T. N. SRINIVASAN, Yale University
DJAVAD SALEHI-ISFAHANI, Virginia Polytechnic Institute & State University
KENNETH BOULDING, University of Colorado

2:30 P.M. RECOMMENDATIONS AND FINDINGS OF AEA COMMISSION ON GRADUATE EDUCATION IN ECONOMICS
Presiding: ROBERT EISNER, Northwestern University
Papers: ANNE O. KRUEGER, Duke University
The Commission's Recommendations
W. LEE HANSEN, University of Wisconsin-Madison
The Commission's Findings
Discussants: MARK BLAUG, University of London
DAVID C. COLANDER, Middlebury College
DANIEL NEWLON, National Science Foundation

2:30 P.M. EMPIRICAL IMPLICATIONS OF MARKET STRUCTURE FOR PRICING BEHAVIOR
Presiding: JOHN LONDREGAN, Carnegie Mellon University
Papers: CRAIG M. NEWMARK, North Carolina State University
Do High Prices Indicate Collusion? A Critical Reiew of Price–Concentration Studies
WILLIAM N. EVANS AND IOANNIS N. KESSIDES, University of Maryland
The Relationship Between Market Share and Price in the Airline Industry
ANDREW S. JOSKOW, GREGORY WERDON, AND RICHARD JOHNSON, U.S. Department of Justice
Empirical Analysis of Entry and Exit in Airline Markets
Discussants: SUGATO BHATTACHARYYA, Carnegie Mellon University

2:30 P.M. THE GOAL OF PRICE STABILITY
Presiding: DAVID J. STOCKTON, Board of Governors of the Federal Reserve System
Papers: LAURENCE BALL, Princeton University
Real Effects of Disinflation with Credible Policy
JACK SELODY, Bank of Canada
The Costs and Benefits of Price Stability: An Empirical Assessment
DAVID LEBOW, JOHN ROBERTS, AND DAVID J. STOCKTON, Board of Governors of the Federal Reserve System
Economic Performance under Price Stability
Discussants: STANLEY FISCHER, World Bank and Massachusetts Institute of Technology
BENNETT MCCALLUM, Carnegie Mellon University
CHARLES SCHULTZE, Brookings Institution

2:30 P.M. ECONOMIC DEVELOPMENTS AND PROSPECTS IN CZECHOSLOVAKIA, YUGOSLAVIA, AND GERMANY (Joint Session with the Association for Comparative Economic Studies)
Presiding: JAN SVEJNAR, University of Pittsburgh
Panel: VLADIMIR DLOUCHY, Deputy Prime Minister of Czechoslovakia
KAREL DYBA, Czechoslovak Academy of Sciences
JAN SVEJNAR, University of Pittsburgh
ZIVKO PREGL, Deputy Prime Minister of Yugoslavia
JANEZ PRASNIKAR, University of Ljubljana, Yugoslavia
HORST SIEBERT, Kiel Institute of the World Economy
IRWIN COLLIER, University of Houston

2:30 P.M. PEACE ECONOMICS: SCOPE, NATURE AND FUTURE DIRECTIONS (Joint Session with Economists Against the Arms Race)
Presiding: SOLOMON W. POLACHEK, State University of New York-Binghamton
Papers: WALTER ISARD, Cornell University, AND CHARLES H. ANDERTON, Holy Cross College
A Survey of Peace Economics and Its Future
Discussants: MARTIN C. MCGUIRE, University of Maryland
JACK HIRSHLEIFER, University of California-Los Angeles
MURRAY WOLFSON, California State University-Fullerton
MANCUR OLSON, University of Maryland

2:30 P.M. HOW ROBUST IS THE INTERNATIONAL FINANCIAL SYSTEM (Joint Session with the International Trade and Finance Association)
Presiding: CATHERINE MANN, Federal Reserve System
Papers: RICHARD J. HERRING, University of Pennsylvania
Dogs that Didn't Bark: What Do They Tell Us about Systematic Risk?
RONALD I. MCKINNON, Stanford University
Exchange-Risk and Interest Rate Volatility Since Bretton Woods
HYMAN P. MINSKY, Jerome Levy Economics Institute
Fragility and Resilience of the International Financial System
H. PETER GRAY, Rensselaer Polytechnic Institute
A Clear and Present Danger?

4:45 P.M. PRESIDENTIAL ADDRESS AND BUSINESS MEETING
Presiding: THOMAS C. SCHELLING, University of Maryland
Speaker: GERARD DEBREU, University of California-Berkeley

Sunday, December 30, 1990

8:00 A.M. MOBILITY, WAGES, AND GENDER
Presiding: KATHARINE ABRAHAM, University of Maryland
Papers: LISA M. LYNCH, Massachusetts Institute of Technology
The Role of Off-the-Job versus On-the-Job Training for Wage Growth and Mobility of Women Workers
THERESA J. DEVINE, Pennsylvania State University
Gender Differences in Job Exit Behavior: An Empirical Analysis Using SIPP
CONSTANCE RHIND, Congressional Budget Office
Retirement in the Dual Worker Family
Discussants: MARK GRITZ, University of Washington
KATHRYN SHAW, Carnegie Mellon University
JANE SJOGREN, Simmons College

8:00 A.M. OPTIMAL PRICING AND ADJUSTMENT COSTS
Presiding: DAVID I. ROSENBAUM, University of Nebraska-Lincoln
Papers: JOHN A. CARLSON, Purdue University
Customer Sensitivity to Relative Prices and Frequency of Price Changes
MENG-HUA YE, George Washington University, AND DAVID I. ROSENBAUM, University of Nebraska-Lincoln
Optimal (s,S) Pricing Boundaries Under Stochastic Inflation Rates
JERZY D. KONIECZY, University of Western Ontario
Variable Price Adjustment Costs
Discussants: KENNETH D. BOYER, Michigan State University
LAWRENCE W. MARTIN, Michigan State University
JULIO ROTEMBERG, Massachusetts Institute of Technology

8:00 A.M. THE ECONOMIC IMPACT OF IMMIGRATION
Presiding: DANIEL HAMERMESH, Michigan State University
Papers: GEORGE J. BORJAS, University of California-Santa Barbara
Immigrants in the U.S. Labor Market: 1940–1980
DAVID CARD, Princeton University
Immigration and the Earnings of Natives
ROBERT J. LALONDE AND ROBERT H. TOPEL, University of Chicago
The Assimilation of Immigrants
Discussants: LARRY KATZ, Harvard University
STEPHEN J. TREJO, University of California-Santa Barbara
JOHN ABOWD, Cornell University

8:00 A.M. AFRICA'S IMMISERIZATION: WHICH WAY FORWARD?
Presiding: E. WAYNE NAFZIGER, Kansas State University
Papers: ELLIOT BERG, Development Alternatives, Inc.
Structural Adjustment and Its Critics: Are We On the Right Path?
FRANCES STEWART, Oxford University
Adjustment with a Human Face
E. WAYNE NAFZIGER, Kansas State University, AND HOWARD STEIN, Roosevelt University
Africa's Economic Development: A Critique of the World Bank and the Lagos Plan of Action
Discussants: ROLF VAN DER HOEVEN, UNICEF
BADIUL A. MAJUMDAR, Washington State University
FANTU CHERU, American University

8:00 A.M. THE DISTRIBUTION OF THE BENEFITS OF NONPROFIT INSTITUTIONS
Presiding: CHARLES T. CLOTFELTER, Duke University
Papers: DICK NETZER, New York University
Arts and Cultural Institutions
JEFF BIDDLE, Michigan State University
Religious Congregations
SANDRA BAUM, Skidmore College, AND SAUL SCHWARTZ, Tufts University
Educational Institutions

Discussants: HELEN LADD, Duke University
EMMETT CARSON, Ford Foundation
ESTELLE JAMES, State University of New York-Stony Brook

8:00 A.M. BEHAVIORAL FINANCE
Presiding: COLIN CAMERER, University of Pennsylvania
Papers: LAURIE SIMON BAGWELL, Northwestern University
Dutch Auction Repurchases: An Analysis of Shareholder Heterogeneity
KENNETH FRENCH, University of Chicago, AND JAMES POTERBA, Massachusetts Institute of Technology
The World Cross-Border Investment Puzzle
JOSEF LAKONISHOK, University of Illinois-Urbana-Champaign, ANDREI SHLEIFER, University of Chicago, RICHARD THALER, Cornell University, AND ROBERT VISHNY, University of Chicago
Do Portfolio Managers Window Dress?
RICHARD ZECKHAUSER, JAY PATEL, AND DARRYLL HENDRICKS, Harvard University
What Investment Flows Reveal About Investory Behavior
Discussants: SUSAN COLLINS, Council of Economic Advisors
HOWARD KUNREUTHER, University of Pennsylvania
COLIN CAMERER, University of Pennsylvania

8:00 A.M. DISCRIMINATION
Presiding: HARRY HOLZER, Michigan State University
Papers: BARRY T. HIRSCH, Florida State University, AND EDWARD J. SCHUMACHER, R. L. Banks & Associates, Inc.
Labor Earnings, Discrimination, and the Racial Composition of Jobs
RONALD D'AMICO, SRI International, AND NAN L. MAXWELL, California State University-Hayward
Employment During the School-to-Work Transition: An Explanation for Subsequent Black–White Wage Differentials and Bifurcation of Black Income
MICHAEL A. LEEDS, Haverford College
Determinants of Success in the AEA Summer Minority Program
Discussants: WAYNE VROMAN, Urban Institute
CHARLES C. BROWN, University of Michigan
RHONDA WILLIAMS, University of Maryland

8:00 A.M. LIFE-CYCLE ANALYSIS OF TAX INCIDENCE
Presiding: ROSEMARY MARCUSS, Congressional Budget Office
Papers: DIANA FURCHTGOTT-ROTH, American Petroleum Institute
The Measurement of Regressivity: The Case of the Motor Fuels Tax
ANDREW B. LYON AND ROBERT M. SCHWAB, University of Maryland
Estimated Tax Incidence in a Life-Cycle Framework
DON FULLERTON, University of Virginia, AND DIANE LIM ROGERS, Pennsylvania State University
A General Equilibrium Study of Lifetime Tax Incidence
Discussants: JANET FURMAN SPEYRER, University of New Orleans
ROBERT LUCKE, Price Waterhouse
JOHN HAKKEN, Congressional Budget Office

8:00 A.M. ECONOMY IN TRANSITION: ISSUES AND POLICIES (Joint Session with the Chinese Economic Association in North America)
Presiding: BEN-CHIEH LIU, Chicago State University
Papers: CLIFF J. HUANG, Vanderbilt University, AND JIN-TAN LIU, Academia Sinica, R.O.C.
Manufacturing Production Frontier and Efficiency: A Case Study of Electronic Industry in Taiwan
HENRY WAN, Cornell University
Technology Transfer and New Growth Theory
PING WANG, Pennsylvania State University, MAW-LIN LEE, University of Missouri, AND BEN-CHIEH LIU, Chicago State University
Growth Versus Equity: A Comparative Study for Taiwan and South Korea
GARY H. JEFFERSON, Brandeis University, AND XU WENYI, Wuhan University and Brandeis University
The Convergence of Factor Returns in Transitional Economies: The Case of Chinese Industry
Discussants: GALE JOHNSON, University of Chicago
GREGORY C. CHOW, Princeton University
ROBERT EISNER, Northwestern University
ANTHONY Y. C. KOO, Michigan State University

8:00 A.M. ECONOMETRIC STUDIES OF HOSPITAL FACTOR COSTS (Joint Session with the Health Economics Research Organization)
Presiding: DONALD E. YETT, University of Southern California
Papers: MICHAEL GROSSMAN, City University of New York Graduate School and National Bureau of Economic Research, FRED GOLDMAN, SUSAN W. NESBITT, New School for Social Research and National Bureau of Economic Research, AND PAMELA MOBILIA, Brooklyn College, City University of New York and National Bureau of Economic Research
Determinants of Interest Rates on Hospital Bonds
JOHN A. RIZZO, Agency for Health Care Policy and Research
Investment and Capital Costs in Proprietary and Not-for-Profit Hospitals
JACK HADLEY, Georgetown University, AND STEPHEN ZUCKERMAN, Urban Institute
Dynamic Estimates of Multiproduct Hospital Cost Functions: 1980–1986
Discussants: FRANK SLOAN, Vanderbilt University
GERARD J. WEDIG, University of Pennsylvania

8:00 A.M. COMPARATIVE ECONOMIC POLICIES (Joint Session with the Society for Policy Modeling)
Presiding: DOMINICK SALVATORE, Fordham University
Papers: VITTORIO CORBO, World Bank
Economic Policies in Latin America
DEENA KHATKHATE, World Development
Economic Policies in India
ADEBAYO ADEDEJI, Economic Commission for Africa
Economic Policies in Africa
Discussants: JOHN M. PAGE, World Bank
ANDARU RAY, World Bank
STEPHEN O'BRIAN, World Bank

10:15 A.M. LEARNING AND ADAPTIVE ECONOMIC BEHAVIOR
Presiding: KENNETH J. ARROW, Stanford University
Papers: W. BRIAN ARTHUR, Stanford University
On Calibrated Rationality
RICHARD J. HERRNSTEIN, Harvard University
Experiments on Stable Suboptimality in Individual Behavior
JOHN HOLLAND, University of Michigan, AND JOHN MILLER, Santa Fe Institute
Just Don't Eat the Typewriter: Artificial Adaptive Agents in Economic Theory
Discussants: KEN BINMORE, University of Michigan
DREW FUDENBERG, Massachusetts Institute of Technology
JOHN GEANOKOPLOS, Yale University

10:15 A.M. WOMEN, WORK, AND CHILDREN IN CANADA AND THE UNITED STATES
Presiding: WILLIAM T. ALPERT, William H. Donner Foundation and University of Connecticut
Papers: ARLEEN LEIBOWITZ AND LINDA WAITE, Rand Corporation
Employment Behavior Surrounding the First Birth
SOLOMON POLACHEK, State University of New York-Binghamton
Effects of Employment Interruptions on Female Wage Rates
ALICE NAKAMURA AND MASAO NAKAMURA, University of Alberta
Effects of Children on Female Earnings: Employment Behavior and Wage Effects in an Intertemporal Context
Discussants: MARTIN BROWNING, McMaster University
T. PAUL SCHULTZ, Yale University

10:15 A.M. TEENAGE PREGNANCY AND WELFARE DEPENDENCY
Presiding: REBECCA BLANK, Northwestern University
Papers: WILLIAM N. EVANS, WALLACE OATES, AND ROBERT SCHWAB, University of Maryland
Community Characteristics and Teenage Pregnancy
GEORGE CAVE AND DANIEL FRIEDLANDER, Manpower Demonstration Research Corporation
How Do Welfare Employment Programs Achieve Their Impacts? Evidence from Cross-Program Comparisons
SHARON LONG, Urban Institute
Analysis of Welfare Dependency: The Dynamics of Participation in the AFDC and Food Stamp Programs
Discussants: ROBERT PLOTNICK, University of Washington
LAURIE BASSI, Georgetown University
JUNE O'NEILL, Baruch College

10:15 A.M. The Logic of Collective Action: Twenty-Five Years Later
Presiding: Burton Weisbrod, University of Wisconsin-Madison and Northwestern University
Papers: Todd Sandler, Iowa State University
The Logic of Collective Action: A Retrospective Look
Martin McGuire, University of Maryland
Empirical Methods and Collective Action
Mancur Olson, University of Maryland
Unanswered Questions
Discussants: John Tschirhart, University of Wyoming
Robert Tollison, George Mason University
Dwight Lee, University of Georgia

10:15 A.M. Exchange Rate Policy
Presiding: Kathryn Dominguez, Harvard University
Papers: Hali J. Edison, Board of Governors of the Federal Reserve System, and Graciela Kaminsky, University of California-San Diego
Target Zones and Exchange Rate Variability
Vittorio Grilli and Nouriel Roubini, Yale University
Liquidity, Exchange Rates, and the International Transmission of Monetary Policy
Paul R. Krugman, Massachusetts Institute of Technology
Speculative Attack and Regime Collapse
Kathryn Dominguez, Harvard University, and Jeffrey A. Frankel, University of California-Berkeley
Exchange Rate Policy Announcements—When Do They Matter?
Discussants: Robert Cumby, New York University
William H. Branson, Princeton University
Kenneth A. Froot, Massachusetts Institute of Technology
Rudiger Dornbusch, Massachusetts Institute of Technology

10:15 A.M. Patterns of Faculty Retirement
Presiding: Albert Rees, Princeton University
Papers: G. Gregory Lozier and Michael Dooris, Pennsylvania State University
Projecting Faculty Retirement: Factors Influencing Individual Decisions
Sharon P. Smith, Princeton University
Ending Mandatory Retirement in the Arts and Sciences
Alan L. Gustman, Dartmouth College, and Thomas Steinmeier, Texas Technological University
The Effects of Pensions and Retirement Policies on Retirement in Higher Education
Discussants: W. Lee Hansen, University of Wisconsin-Madison
Richard Burkhauser, Syracuse University

10:15 AM. Economics of Drugs
Presiding: Paul J. Taubman, University of Pennsylvania and National Bureau of Economic Research
Papers: Gary Becker, University of Chicago
Legalization of Drugs
Jeffrey Miron, National Bureau of Economic Research
The Effects of Drug Criminalization on Drug Use: Some Cross-Regime Evidence
Robin Sickles, Rice University and National Bureau of Economic Research, and Paul J. Taubman, University of Pennsylvania and National Bureau of Economic Research
Who Uses Illegal Drugs?
Discussants: Claudia Goldin, Harvard University
Mark Kleiman, Harvard University
Peter Reuter, Rand Corporation

10:15 A.M. Seminars on Recent Topics and Techniques*
Speaker: John W. Campbell, Princeton University
New Developments in the Analysis of Stock Markets

10:15 A.M. Labor Market Institutions and Outcomes
Presiding: Lisa M. Lynch, Massachusetts Institute of Technology
Papers: Todd L. Idson and Philip K. Robins, University of Miami
Labor Turnover and Worker Mobility in Small and Large Firms
Takao Kato and Mark Rockel, Colgate University
Managerial Internal Labor Markets and Executive Compensation: A Comparison of Japan and the United States

Janet Currie, University of California-Los Angeles, and Sheena McConnell, Mathematical Policy Research
The Right to Strike versus Compulsory Arbitration in the Public Sector: Effects on Wages and Dispute Costs
Linda C. Calvalluzzo, Union College, and James M. MacDonald, Rensselaer Polytechnic Institute
The Effects of Rail Deregulation on Labor
Discussants: James Rebitzer, Massachusetts Institute of Technology
Sherry Glied, Columbia University
Nancy Rose, Massachusetts Institute of Technology

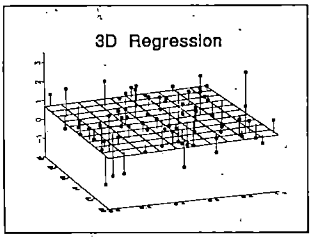
10:15 A.M. When and Why Do Institutions Really Matter? New Evidence from Economic History (Joint Session with the Economic History Association)
Presiding: Jeremy Atack, University of Illinois-Urbana-Champaign
Papers: Michael Bordo and Eugene White, Rutgers University-New Brunswick
A Tale of Two Currencies: British and French Finances During the Napoleonic Wars
William Darity, University of North Carolina-Chapel Hill
Technical Change in Agriculture, Population Growth, and Theories of Economic History
Elizabeth Hoffman and Gary Libecap, University of Arizona
The History and Functioning of a Government-Sponsored Cartel: Agricultural Marketing Orders for Citrus
Discussants: Larry D. Neal, University of Illinois-Urbana-Champaign
Gregory Clark, University of Michigan
Thomas S. Ulen, University of Illinois-Urbana-Champaign

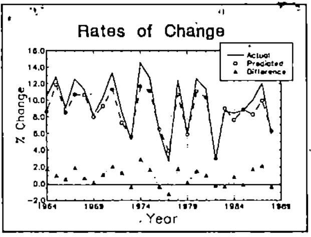
10:25 A.M. Reregulation for the Promotion of Competition? (Joint Session with the Transportation and Public Utilities Group)
Presiding: Benjamin J. Allen, Iowa State University
Papers: Rodney Stevenson, University of Wisconsin-Madison
Regulation, Sharing Rules, and the Quality of Competition
Robert Windle and Martin Dressner, University of Maryland
Competition at "Competitive" Hubs in the U.S. Domestic Air Transport Industry
Edward Morash and George Waggenheim, Michigan State University
The Impact of Federal Deregulation on State Regulatory Attempts
Discussants: Denis Breen, Federal Trade Commission
Joe Kerkvliet, Oregon State University

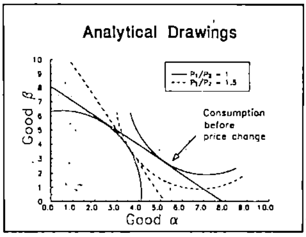
Note: Although the following sessions will be included in the final program, it was not possible at press time to schedule day and time assignments because of incomplete information:

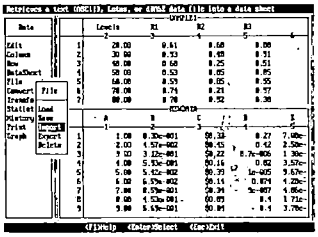
Economic Reform in Poland and Hungary
Applications of Contract Design
What Remains of Comparative Advantage?
Intertemporal Choice
Innovation and Transaction-Cost Economics
Western Europe, Eastern Europe, and the World Economy
Topics in Macroeconomics
Consumption, Saving, and Interest Rates
Output, Employment, Prices
The Economics of Health Care
Teenage Pregnancy and Welfare Dependency
Mitigation of and Adaptation to Global Environmental Change
Protection of the Individual
Political Economy
History of Economic Thought
Greenhouse Warming
Demand Uncertainty, Inventory Investment, and Seat Belts
Asian Economic Regimes
Economics and Public Policy
Information
Bidding and Negotiation
Depository Institutions
The Political Economy and Monetary Policy
Economic Effects of Government Credit Market Policies
Tropical Forest Protection (joint with AERE)
Frontiers of Cliometric Research (joint with CS)
ODE Chapter Advisers and Regional Directors Session (joint with ODE)

*The American Economic Association will offer a new series of seminars on recent topics and techniques at the 1990 annual meeting. At these seminars, a member of the Association will lecture on a recently developed technique or research area. Reading lists will be made available so that participants can further develop their interest. The aim of these seminars is to reduce the time between development of new knowledge and its availability in textbooks or survey articles. Seminars will last approximately two hours.



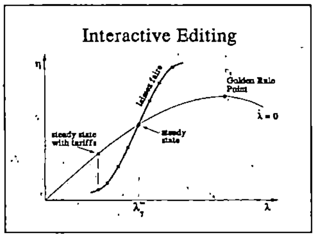


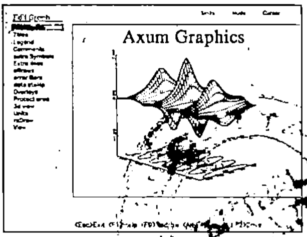



Retrieves a text (ASCII), Lotus, or dBASE data file into a data sheet

	Levels	X1	X2	X3
1	20.00	0.61	0.60	0.00
2	30.00	0.53	0.49	0.51
3	40.00	0.68	0.25	0.51
4	50.00	0.53	0.65	0.85
5	60.00	0.59	0.85	0.55
6	70.00	0.74	0.21	0.57
7	80.00	0.70	0.52	0.38

	A	B	C	D	E
1	1.00	0.30e-001	90.32	0.27	7.90e-
2	2.00	4.57e-002	90.45	0.42	2.50e-
3	3.00	3.12e-001	90.22	8.7e-006	1.30e-
4	4.00	5.93e-001	90.16	0.02	3.57e-
5	5.00	5.42e-002	90.39	1e-005	9.67e-
6	6.00	6.59e-002	90.14	0.074	4.20e-
7	7.00	8.59e-001	90.34	5e-007	4.86e-
8	8.00	4.53e-001	90.65	0.4	1.71e-
9	9.00	5.69e-001	90.04	0.4	3.70e-





Amazing graphics made simple!

Here is what Axum users are saying:

"Fantastic..." "In a matter of minutes I took my Lotus data, created a graph, and used it in WordPerfect..." "Axum is exactly what I've been waiting for." "I created publication-quality graphs without even using the manual..." "...other graphics packages are Stone Age compared withAxum!"

Introducing Axum...

Axum, a new technical graphics and data analysis program for PC's, combines the ease-of-use of the best-selling business graphics packages with the advanced capabilities of expensive mainframe technical graphics programs.

Even if you have never used a graphics program before, with Axum you will be analyzing and graphing your data within minutes.

Publication-quality technical graphs

Take advantage of Axum's extraordinary display capabilities.

- Produce publication-quality, color graphs
- Choose from numerous graph types: 2D, 3D, Contour & more
- Choose from 21 fonts including Greek and scientific; use unlimited super and subscripts
- Graph arbitrary 2D and 3D functions
- Interactively create and customize
- Place multiple graphs on a page: rotate, shrink, or expand
- Automatic regression plots
- Scale plots to multiple y axes and logarithmic axes
- Change the viewpoint of your 3D graph
- Features ultra-high-resolution output from GraphiC™ by Scientific Endeavors

State-of-the-art data editor

Use Axum to perform both simple and advanced data analysis with ease.

- Easily import, edit and transform data
- Explore and analyze your data quickly and easily using multiple, interrelated "data sheets"
- Use unlimited-sized data sets
- Do statistics, curve fitting, and smoothing
- Create macros using Axum's built-in advanced programming language
- Choose from over 100 functions and operators for manipulating and transforming your data

Works with the software you already have

- Use your graphs in WordPerfect, Word, Ventura, PageMaker
- Export to GEM, Lotus PIC, HPGL, Tektronix, and Encapsulated PostScript files
- Import ASCII, Lotus, and dBase files
- Use virtually any monitor, printer (including PostScript) and plotter

60 day money back guarantee

Try Axum without risk. If you are not completely satisfied, return it within 60 days for a refund.

Axum™

Call toll free for FREE brochure

1-800-548-5653 ext. 029

TRIMETRIX

444 NE Ravenna Blvd Suite 210-AE
Seattle, WA 98115
TEL. WA 206-527-1801
FAX 206-522-9159
For the IBM PC, XT, AT, PS/2 and compatibles